Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting

ANONYMOUS AUTHOR(S)

Counterfactual explanations (CFEs) highlight what changes to a model's input would have changed its prediction in a particular way. CFEs have gained considerable traction as a psychologically grounded solution for explainable artificial intelligence (XAI). Recent innovations introduce the notion of computational plausibility for automatically generated CFEs, enhancing their robustness by exclusively creating plausible explanations. However, practical benefits of such a constraint on user experience and behavior is yet unclear. In this study, we evaluate objective and subjective usability of computationally plausible CFEs in an iterative learning design targeting novice users. We rely on a novel, game-like experimental design, revolving around an abstract scenario. Our results show that novice users actually benefit less from receiving computationally plausible rather than closest CFEs that produce minimal changes leading to the desired outcome. Responses in a post-game survey reveal no differences in terms of subjective user experience between both groups. Following the view of psychological plausibility as comparative similarity, this may be explained by the fact that users in the closest condition experience their CFEs as more psychologically plausible than the computationally plausible counterpart. In sum, our work highlights a little-considered divergence of definitions of computational plausibility and psychological plausibility, critically confirming the need to incorporate human behavior, preferences and mental models already at the design stages of XAI approaches.

Additional Key Words and Phrases: XAI, counterfactual explanations, quantitative user evaluation, Data and Algorithm Evaluation, Human Factors

ACM Reference Format:
Anonymous Author(s). 2022. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. In . ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Explaining one's behavior to another is a critical element in human social interaction. A person depends on explanations to improve their understanding, ultimately building a stable mental model as basis for prediction and control [31]. The need to effectively explain not just human action, but also the behavior of automated systems and their underlying machine learning (ML) models, has received increasing attention in recent years. This development gave rise to the increasing interest in explainable artificial intelligence (XAI) as a research field. Consequently, the XAI community has seen a veritable surge of technical accounts on how to realize explainability for ML [30].

Motivated by a seminal review by Miller advocating a user-centered focus on explainability, counterfactual explanations (CFEs) gained particular prominence as a supposedly useful, human-accessible solution [37, 50]. CFEs provide *what-if* feedback to the user, i.e., information on what changes in the input elicit a change of an automated decision (i.e., "if you had worn a mask, you would not have gotten ill"). However, the emerging body of work on CFEs, and

Manuscript submitted to ACM

explainability of ML models more generally, shows an alarming tendency to take the quality of the suggested explanation modes at face value [24, 52]. A recent review of counterfactual XAI studies reveals that only one on three studies concern themselves with user-based evaluations, often with limitations concerning statistical power and reproducibility [37].

The lack of user-based evaluations affects not only assessments of CFEs as such, but more specifically also the evaluation of different conceptualizations for this kind of explanations. The prevailing approach in the current literature is to compare different CFE approaches exclusively in terms of their robustness and theoretical fairness [6, 21, 70], passing over the role of the user as eventual target. Thus, in-depth evaluations of user experiences, elucidating the usability of CFE variants, are yet to be done.

The current work marks a step towards closing this fundamental research gap, focusing on the concept of plausibility. While technical descriptions of plausible CFEs approaches exist [5, 61, 63], no user study to date has directly investigated potential benefits of enforcing an additional plausibility constraint. Thus, we perform a well-powered user study analyzing the performance of novice users when receiving closest CFEs exclusively defined via their proximity to the decision boundary, compared to computationally plausible CFEs as feedback in an iterative learning design [4, 5].

2 COUNTERFACTUAL EXPLANATIONS AS A PSYCHOLOGICALLY GROUNDED SOLUTION FOR XAI

A major challenge for XAI is the lack of a common, straight-forward and universally applicable definition of what constitutes a good explanation. To complicate matters, the effectiveness of an approach may depend on the reason for explaining [1], as well as pre-existing knowledge and experiences of users at the receiving end [66].

In search of truly human-usable explanation modes, the XAI community recognized the need to bridge the gap between psychology and computer science in order to draw inspiration from how humans explain in their daily social interactions [50]. A central insight from classical psychological literature is that human explanations are typically contrastive: They emphasize (explicitly or implicitly) why a specific outcome occurred instead of another [32, 43, 45, 50].

This contrastive nature relates to the more general human tendency to reflect upon past events by generating possible alternatives, i.e., counterfactual thinking [58]. Empirical evidence demonstrates that humans show this *what-if* mentality spontaneously [28], and increasingly when facing negative outcomes or unexpected results [60]. In their functional theory of counterfactual thinking, Roese and Epstude suggest a crucial role of counterfactual thoughts to guide to formation of future intentions, thus regulating subsequent behavior [25, 59]. This evidence is the root for the common supposition in XAI that explanations formulated as counterfactuals are naturally intuitive, easy to understand, and helpful for users, often discounting the need for user evaluations [5, 17, 29, 65].

Decades of philosophical and psychological research has concerned itself with the question of how humans generate counterfactuals. Lewis' seminal work on the topic builds on a theory of possible worlds, postulating that counterfactual statements trigger a comparison between the actual circumstances and a conceivable world in which the counterfactual statement occurred [41]. Embedding this view into a cognitive framework of counterfactual thought, the mental models theory emphasizes the human ability to entertain two parallel representations of reality: The factual conditional, corresponding to the true state of the world, and the concurrent non-factual possibility, temporarily assumed to be true [11, 13, 34, 69]. Insights from neuroimaging support this notion, demonstrating that counterfactual thinking extends mere hypothetical deliberation by recruiting additional representational processes in the brain [38].

When humans generate counterfactuals, they show remarkable regularities in terms of which aspects of the past they reconstruct. Humans tend to modify events that are recent [14, 49], exceptional, while also regarding the optimal counterfactual outcome [23, 35], and controllable events when undoing of fictitious outcomes [26]. Further, authors like to note that humans produce plausible rather than implausible counterfactuals [12, 20].

However, despite being a commonly-used notion in psychology, plausibility is difficult to define precisely. Variable interpretations of what constitutes a plausible counterfactual exist, referring to different partially overlapping concepts. Kahneman and Tversky refer to hypothetical events as plausible if they are easy to imagine [35]. Lewis supposes that plausible counterfactuals come from worlds that are minimally different from reality [41]. Building up on this idea of comparative similarity, empirical research shows perceived plausibility of a counterfactual event to be proportional to the perceived similarity between said counterfactual and the factual state [19, 64].

In addition to such a similarity-based definition, plausibility is often used synonymously with concepts of likeliness or probability [18, 55]. De Brigard et al. demonstrate that manipulations of counterfactual plausibility in terms of their likeliness changes their neural representation [18]. Their findings may indicate greater affective evaluation for counterfactuals that carry greater subjective likelihood, and thus, plausibility. In their plausibility analysis model, Connell and Keane expand on the idea of plausibility as probability and highlight the pivotal role of pre-existing domain knowledge, postulating that a scenario may only be plausible if it fits well to prior knowledge [16].

Thus, while it is difficult to pinpoint exactly what makes a counterfactual psychologically plausible, we may recognize pivotal roles of concepts like comparative similarity and probability. Following the user-centered focus on explainability proposed by Miller [50], incorporating these concepts would be an important step towards automatic generation of plausible, and thus more human-friendly and usable, CFE.

3 COMPUTATION OF CFES AND PLAUSIBLE CFES

Wachter et al. introduce a CFE $\vec{x}_{\text{cf}} \in \mathbb{R}^d$ of an ML model $h : \mathbb{R}^d \rightarrow \mathcal{Y}$ as an optimization problem [68] :

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min} \; \ell\left(h(\vec{x}_{\text{cf}}), y'\right) + C \cdot \theta(\vec{x}_{\text{cf}}, \vec{x}) \tag{1}$$

where $\vec{x} \in \mathbb{R}^d$ denotes the original input, the regularization $\theta(\cdot)$ penalizes deviations from the original input $\vec{x}$ (weighted by a regularization strength $C > 0$), $y' \in \mathcal{Y}$ denotes the requested output/behavior of the model $h(\cdot)$ under the counterfactual $\vec{x}_{\text{cf}}$, and $\ell(\cdot)$ denotes a loss function penalizing deviations from the requested prediction.

Thus, computing CFEs translates to finding minimal perturbations to a model's input that alter the final prediction to a desired outcome. Given the regularization term $\theta(\cdot)$, generated CFEs based on this definition remain as close to the original input $\vec{x}$ as possible. Consequently, we will refer to them as *closest CFEs* for the remainder of this work. As one of the first approaches to model CFEs for classical ML, Eq. (1) is the forerunner of more powerful, model specific variations, as well as many methods for solving these optimization problems [3, 36, 67]. However, it is important to note that *closest CFEs* do not necessarily yield plausible or even realistic counterfactuals. As a matter of fact, *closest CFEs* may look like adversarials, introducing only slight changes in the input that go unnoticed by a human observer despite altering the model's output [53]. Whether a computed *closest CFEs* corresponds to such an adversarial depends on the model, loss function and regularization, diminishing their suitability as explanation technique [40]. Expanding the original definition in Eq. (1) by an additional plausibility constraint circumvents these issues:

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min} \; \ell\left(h(\vec{x}_{\text{cf}}), y'\right) + C \cdot \theta(\vec{x}_{\text{cf}}, \vec{x}) \qquad \text{s.t. } \vec{x}_{\text{cf}} \in \mathcal{P} \tag{2}$$

where $\mathcal{P}$ denotes the set of all plausible CFEs.

To distinguish these explanations from their psychologically plausible counterparts discussed in the previous section, we will henceforth refer to these solutions as *computationally plausible CFEs*. Similar to the modeling of *closest CFEs* in Eq. (1), different realizations of *computational plausibility* have been proposed [4, 46, 56]. One particular instance

are density based approaches [4] that restrict a counterfactual $\vec{x}_{\text{cf}}$ to regions of high density. The respective density function may for instance be estimated from the training data. In the current work, we follow an alternative approach when providing *computationally plausible CFEs* and limit the set of possible counterfactuals to the training data as a representative set of feasible examples [56].

4 DO NOVICE USERS PROFIT FROM COMPUTATIONAL PLAUSIBILITY IN AN ABSTRACT DOMAIN?

The guiding question of the current work is whether *computationally plausible CFEs* have an advantage over *closest CFEs* in helping users to learn from an ML model. To assess this question, we rely on an interactive iterative learning task, where users repeatedly choose input values for an ML model. At regular intervals, they receive either *computationally plausible* or *closest CFEs*, highlighting how changes in the user's previous input would have lead to better results. The main advantage of this approach is that the interplay between repeated user action and corrective feedback enables us to assess user understanding at each stage of the process objectively through task performance.

We find it conceivable that implementing a plausibility constraint indeed improves user performance. Specifically, we assume that repeated exposure to items representative of the training set enables humans to build a more accurate mental model of the underlying data distribution. To obtain general insights about the usability of different types of CFEs as such, we recruited novice users and designed the task around an abstract scenario. This approach has the additional advantage to mitigate any difference in domain knowledge and possible misconceptions about the task setting, potentially confounding task performance [66]. Thus, we formulated the following three hypotheses.

Hypothesis 1. We expect *computationally plausible CFEs* to be more helpful to users tasked to discover unknown relationships in data than *closest* ones, both objectively and subjectively. Specifically, we anticipate that participants in the plausible condition a) show greater learning success, b) become more automatic and thus quicker in the task, and c) are able to explicitly identify relevant and irrelevant input features.

Hypothesis 2. We expect a group difference in terms of subjective understanding. We predict that users will differ in how far they find CFEs useful, and in how far they can utilize them, with an advantage of *computationally plausible CFEs*. Furthermore, we posit that users imagine *computationally plausible CFEs* to be more helpful for other users.

Hypothesis 3. We evaluate users' understanding of the explanations themselves, their need for support to understand, and their evaluation of timing and efficacy of CFE presentation. As structure and presentation mode of CFEs is kept constant across conditions, we expect not to find any differences. This analysis tests the comparability of conditions, a key feature in any experimental user design.

Finally, we do not formulate a prediction whether groups will differ in uncovering inconsistencies in the explanations presented. This will be investigated in an additional exploratory analysis.

5 EXPERIMENTAL DESIGN

To assess *Hypotheses 1–3*, we use a novel iterative learning design revolving around an abstract scenario. Figure 1 conveys the overall two-part structure of the study.

5.1 The Plausible Alien Zoo Scenario

We developed a game-like experimental design on a web-based interface to provide global access for users from diverse backgrounds, facilitating large-scale participant recruitment.

In the Alien Zoo scenario, participants imagine themselves as zookeepers for aliens. To feed to the aliens, participants may choose from different plants. However, it is not clear what plants make up a nutritious diet. Thus, participants

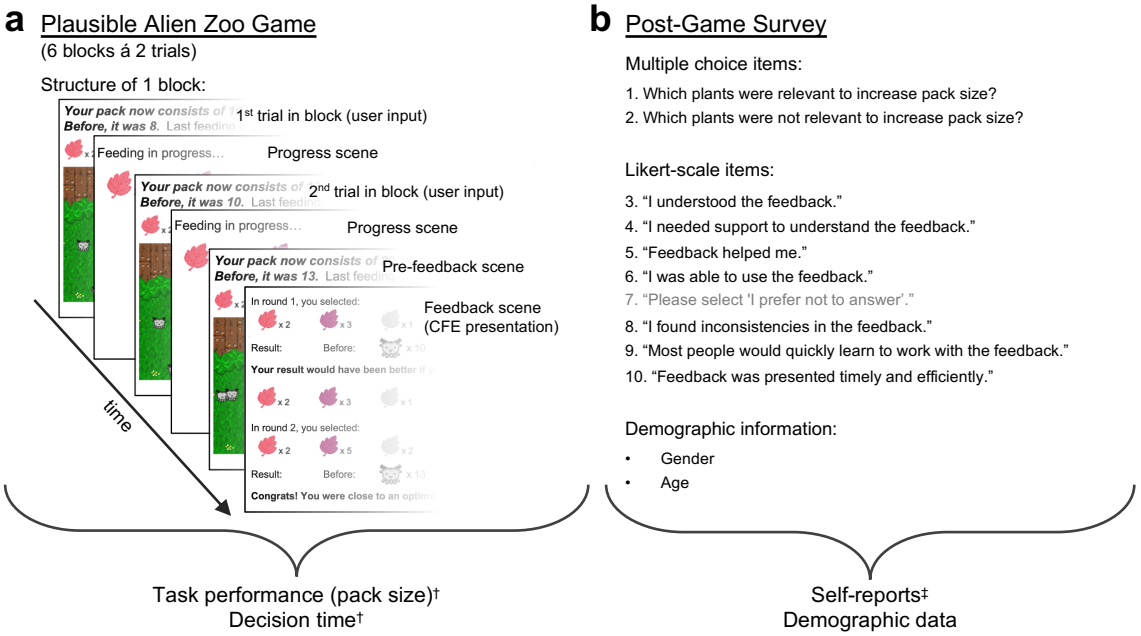

Fig. 1. General overview of study procedure. (**a**) The plausible Alien Zoo game is an iterative design with blocks containing two trials calling for user input, and finishing with a feedback scene that provides either *closest* or *computational plausible CFEs* computed on user's previous input. (**b**) After 6 blocks, users enter a post-game survey collecting self-report information from participants. Likert-scale items adapted from [33]. Catch item marked in light gray. The lower part of both subfigures shows measures evaluates from respective parts; † objective measure, ‡ subjective measure.

need to find how to best feed the aliens. Participants go through several feeding cycles, choosing a combination of plants. After each cycle, the pack of aliens either decreases (given a bad combination of plants) or increases (given a good combination). In regular intervals, participants receive a summary of their past choices, together with feedback on what choice would have led to a better result (i.e. a CFE).

Assessing performance of real users in an abstract task setting, this use case corresponds to a human grounded evaluation [24]. Further, our setting falls under the "explaining to discover" category for explainability defined by Adadi and Berrada, investigating whether providing CFEs to novice users improves their understanding of relationships in a yet unknown dataset [1].

5.2 Post-Game Survey

A post-game survey collects self-report information from participants. Besides explicitly asking participants to point out which plants were relevant and irrelevant for the task, we use an adapted version of the System Usability Scale [33], designed to measure the quality of explanations elicited by an explainable ML system. Participants answer a series of Likert-scale items, assessing how users feel about using our system with a focus on perceived understandibility and usability of CFEs. The survey closes with asking for participants' gender and age as potential confounding variables. Figure 1b gives a complete overview of all items in the survey part, in the order participants encounter them.

5.3 Constructs and Measurements

We measure understanding and usability of explanations in terms of two objective behavioral variables and several subjective self-reports (Figure 1 bottom).

Regarding task performance, we assess the development of pack size in the Alien Zoo game over trials. This value indicates 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.

Second, we measure time needed to reach a feeding decision over trials (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 [44].

We acquire self-reports via the post-game survey, assessing different aspects of participant's system understanding. The first two survey items ask users to identify plants they think are relevant and irrelevant for task success. Replies from these items allow us to measure to which extent users in different groups formed explicit knowledge of the underlying data structure. Further, users indicate in how far they find the explanations 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 understanding.

Finally, three self-report measures check for potential confounds. These are items that ask users to indicate their understanding of the explanations as such, whether they feel the need for support for understanding, and their evaluation of timing and efficacy of CFE presentation. Given that structure and presentation mode of CFEs is kept constant for both groups, differences would uncover unexpected variation in terms of answer style, a potential confounding variable.

5.4 Implementation, ML Model and Data Set

The back end of the system is written in Python3, using the sklearn package [54] for the ML part. The front end employs the JavaScript-based Phaser 3, an HTML5 game framework[1]. We use a decision tree regression model for predicting the growth rate given the plants selected by the user. Decision trees approximate the data distribution with a collection of if-then-else rules, consecutively splitting the data [62]. We choose a decision tree because computing counterfactuals for this model is fairly simple [3]. Yet, it is powerful enough to model our synthetic data set sufficiently well. The current implementation uses the Gini splitting rule of CART [10], with a maximum tree depth of 4. The decision tree corresponding to the ground truth model is build once in the beginning and remains the same for all users during the study.

We use the code provided by the CEML package [2] for computing CFEs. The underlying data set used for tree building consists of 5 integer features (i.e., the plants used for feeding) and 1 continuous output variable (i.e., the growth rate used as factor for computing the new pack size). We generated this data according to the following scheme: The growth rate scales linearly with plant 2, iff plant 4 has a value of 1 or 2 AND plant 5 is not smaller than 4. Growth rate may take a value between 0 and 2, used as a factor for pack size in the previous round to compute the new pack size. The initial full data set contains all possible plant – growth rate combinations 100 times, yielding 3 276 800 data points. For final model training, we sample a subset of 10 922 data points from this full set to introduce sparsity, thus ensuring that computed *closest* and *computationally plausible CFE* diverge. Note that our implementation prevents pack size from shrinking below 2.

[1] https://phaser.io/

5.5 Participants

The study ran in early November 2021 on Amazon Mechanical Turk (AMT). After performing three pilots with 10 users each to refine the experimental design, we recruited a total of 100 participants for final assessment. A first data quality check revealed corrupted data for four participants due to logging issues. Thus, we acquired four additional data sets. All participants gave informed electronic consent by providing clickwrap agreement prior to participation. All participants received a reward of US$ 4 for participation. The ten best performing users received an additional bonus of US$ 2. Game instructions informed participants about the possibility of a bonus to motivate compliance with the experimental task [7]. The study was approved by the Ethics Committee of the author's home university.

5.6 Experimental Procedure

After accepting the task on AMT, participants are forwarded to our web server hosting the alien zoo game. They first encounter a page informing them about purpose, procedure and expected duration of the study, their right to withdraw, confidentiality and contact details of the primary investigator. Users may decline to participate by closing this window. Otherwise, they indicate their agreement via button press, opening a new page. Unbeknownst to the user, they are randomly assigned to either the *closest* or the *plausible* condition when they indicate agreement.

The succeeding page provides detailed instructions to the game. Specifically, it shows images of the aliens, as well as the selection of plants they may choose to feed from. Written instructions detail that it is possible to choose up to six leaves per plant in whatever combination seems desirable, and that choosing healthy or unhealthy combinations lead to increases or decreases in pack size, respectively. Further instructions emphasize the user's task to maximize the number of aliens, so-called shubs, with the best players qualifying for a monetary bonus. Participants are also informed that they will receive feedback on what choice would have led to a better result after two rounds of feeding. Users may indicate that they are ready to begin the game by clicking a "Start" button at the end of the page. To prevent participants from skipping the instructions, this button appears with a delay of 20s.

Upon hitting "Start", participants encounter a padlock scene where they can make their feeding choice (Figure 2, top left image). The right side of the screen displays leaves from all plant types next to upward/downward arrow buttons. In the first feeding round, the top of the page shows written information that clicking on the upward arrows increases the number of leaves per plant, while clicking the downward arrows has the reverse effect. In each succeeding feeding round, the top of the page shows the current pack size, the pack size in the previous round, and the choice made in the previous round. The page additionally shows a padlock with the current number of animated shubs. Each participant starts of with a pack of 10 aliens. After making their choice, participants continue by clicking a button stating "Feeding time!" in the bottom right corner of the screen.

Upon committing their choice, a progress scene displaying the current choice of plants and three animated aliens is shown. Meanwhile, the underlying ML model uses the user input to generate the new growth rate and pack size, together with either a *closest* or a *computationally plausible CFE*. After 3s, the padlock scene appears again to show the results of their last choice. Following odd trials, the user may make a new selection. After even trials, a single "Get feedback!" button replaces the choice panel on the right-hand side of the screen. Hitting the feedback button forwards a user to an overview scene displaying the feeding choices in the last two runs, the resulting changes in pack size and the counterfactuals that indicate what choices would have led to better results. When users made a choice that led to maximal increase in pack size such that no counterfactual could be computed, they are told that they were close to an optimal solution in that round. Users may move on to the next round by hitting a "Continue!" button appearing after

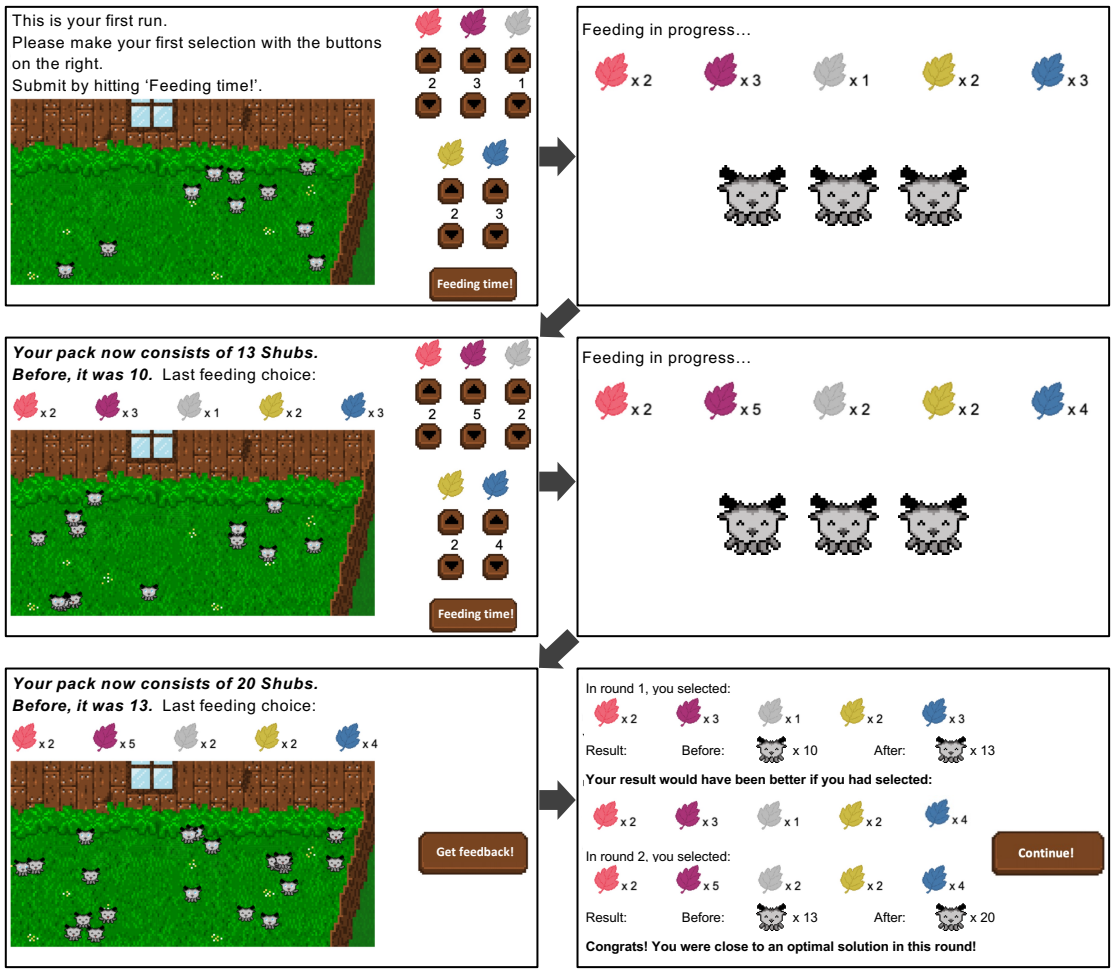

Fig. 2. Exemplary user journey through the first block of the plausible Alien Zoo game. Bold arrows indicate temporal succession of respective scenes. The figure highlights the iterative nature of the game with repeated user input and end-of-block presentation of CFEs. Note that plant counters are set 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.

10s on the right-hand side of the screen. This delay forces users to spend some time with the information to study it. Upon continuing, users make their new choice in a new padlock scene.

The study runs over 12 feeding rounds (trials) with feedback interspersed after each second trial. To ensure attentiveness of users during the game, we included two additional attention trials. After feeding rounds 3 and 7, users face a new page requesting to type in the current number of aliens in their respective packs. Immediate feedback informs participants whether their entry was correct or not, and reminds users to pay close attention to all aspects of the game at any given time. Subsequently, the next progress scene appears and the game continues.

The game part of the study is complete after 12 trials. The experimental procedure concludes with a survey assessing user's explicit knowledge on what plants were and were not relevant for improvement (items 1 and 2), as well as

an adapted version of the System Causability Scale [33] evaluating the subjective quality of explanations. The study closes with two items assessing demographic information on gender and age. The final page thanks users for their participation and provides a unique code to insert in AMT to prove that they completed the study and qualify for payment. Further, participants may choose to visit a debriefing page with full information on study objectives and goals.

On average, participants needed 13m:43s (± 00m:23s SEM) from accepting the HIT on AMT to inserting their unique payment code.

5.7 Statistical Analysis, Sample Size Calculation and Data Quality Measures

We perform all statistical analyses using R-4.1.1 [57], using CFE variant (*closest* or *computationally plausible*) as independent variable. Changes in performance over 12 trials as a measure of learning rate per group (lme4 v.4_1.1-27.1) [8]. In the model testing for differences in terms of user performance, the dependent variable is number of aliens generated. In the assessment of user's reaction time, we use decision time in each trial as dependent variable. The final models include the fixed effects of group, trial number and their interaction. The random-effect structure includes a by-subjects random intercept. Advantages of using this approach include that these models account for correlations of data drawn from the same participant and missing data [22, 51]. The analysis of variance function of the stats package in base R serves to compare model fits. η_p^2 values denote effect sizes (effectsize v.0.5) [9]. Computation of pairwise estimated marginal means follow up significant main effects or interactions, with respective effect sizes reported in terms of Cohen's *d*. All post-hoc analyses reported are Bonferroni corrected to account for multiple comparisons.

We evaluate data gathered from the post-game survey depending on question type. For the first two items assessing user's explicit knowledge of plant relevance, we test data for normality of distributions 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. We also compare user's explicit knowledge of plant relevance to the expected value given random response patterns using the non-parametric one-sample Wilcoxon signed rank test for each group separately, and report Bonferroni corrected results. To analyze group differences of ordinal data from the Likert-style items, we rely on the non-parametric Wilcoxon-Mann-Whitney *U* test. We report effect sizes for all survey data comparisons as *r*.

We performed an a priori sample size estimation based on data obtained from the third pilot. To do so, we set up two linear mixed effects models with pack size and reaction time as described above. For each of these models, we run a simulation-based power analysis for samples of 20–100 participants with fixed effects of group and trial number over 1000 simulations (mixedpower v.0.1.0) [39]. Inspecting the results, we choose to acquire data from 100 participants to reach a power > 80% for at least one combination of trial number and group, while also accounting for potential participant attrition.

As a web-based study, we run the risk that some participants attempt to game the system to collect the reward without providing proper answers. Thus, we implement a number of data quality checks that were planned a priori. We identify speeders based on the decision time, flagging users that spent less than 2s in the padlock scene in 4 or more trials. We flag participants that fail to respond with the correct number of aliens in both attention trials during the game. Furthermore, we included a catch item in the survey (1b, item 7), flagging inattentive users. Finally, we identify straight-lining participants who keep choosing the same plant combination despite not improving in at least two blocks, or answer with only positive or negative valence in the survey. To uphold a high threshold for data quality, we follow a conservative approach of excluding participants that were flagged for at least one of these reasons.

Table 1. Demographic information of participants.

	Before quality assurance measures (N = 100)				After quality assurance measures (N = 74)			
	closest	*plausible*	U value[a]	p value	*closest*	*plausible*	U value[a]	p value
N	50	50	..	..	40	34	..	..
Gender[b]	17f/33m	22f/26m/1nb/1na	1108	.339	13f/27m	18f/15m/1nb	554.4	.116
Age (*Mdn*)[c]	25–34y	25–34y	1234	.950	25–34y	35–44y	712.5	.718

[a] non-parametric Wilcoxon-Mann-Whitney U test
[b] f = female, m = male, nb = non-binary / gender non-conforming, na = no gender information disclosed
[c] *Mdn* = median age band (options: 18-24y, 25-34y, 25-34y, 35-44y, 45-54y, 55-64y, 65y and over)

6 RESULTS

From 100 participants recruited via AMT, we exclude data from participants who qualified as speeders (n = 2), failed both attention trials during the game (n = 5), gave an incorrect response for the catch item in the survey (n = 3), or straight-lined during the game (n = 4) or in the survey (n = 12), leaving data from 74 participants for final analysis (Table 1).

6.1 Do Computationally Plausible CFEs Facilitate Learning?

Hypothesis 1 postulates that users in the *plausible* condition outperform users in the *closest* condition. To statistically assess 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 3a shows the development of average pack size as well as average decision time per group. Strikingly, the data suggests that participants in the *closest*, not the *plausible*, condition performed better. This effect is confirmed by the significant interaction of factors *trial number* and *group* (F(11,792) = 2.119, p = .017, η_p^2 = 0.029) in the corresponding linear mixed effects model. The follow-up analysis reveals significant differences between groups in trial 11 (t(472) = 4.040, p = .012, d = 0.693) and trial 12 (t(472) = 2.530, p <.001, d = 1.101). Additionally, there is a highly significant main effect of trial number (F(11,792) = 7.585, p <.001, η_p^2 = 0.095), but no significant main effect of group (F(11,72) = 2.586, p = .112, η_p^2 = 0.035).

Participants in both groups showed a marked decrease in decision time over the curse of the study, already apparent after the very first trial (Figure 3b). The significant main effect of factor *trial number* (F(11,792) = 14.818, p <.001, η_p^2 = 0.171) confirms this observation. Corresponding post-hoc analyses show significant differences between trial 1 and all other trials (all t(792) >5.900, p <.001, d >1.200), between trial 3 and 4 (t(792) = 3.765, p = .012, d = 0.621), and between trials 4 and 5 (t(792) = 3.395, p = .048, d = 0.560). Neither the main effect of factor *group* (F(11,72) = 0.235, p = .630, η_p^2 = 0.003), nor the interaction between factors *trial number* and *group* (F(11,792) = 0.897, p = .543, η_p^2 = 0.012) reach significance.

In terms of mean number of matches between user judgments of plant relevance for task success and the ground truth, users in both groups performed comparably both for relevant (*closest*: mean number of matches = 2.850 ± 0.198 *SE*; *plausible*: mean number of matches = 3.206 ± 0.178 *SE*; U = 781, p = .255, r = .054) and irrelevant plants (closest: mean number of matches = 3.125 ± 0.157 *SE*), plausible: mean number of matches = 3.177 ± 0.217 *SE*; U = 721.5, p = .643, r = .054).

While groups did not differ in terms of matches between user judgments of plant relevance, we find significant differences between mean group responses compared to the expected value given random responses (i.e., expected mean number of matches = 2.500): Users in the *computationally plausible* group have significantly more matches than

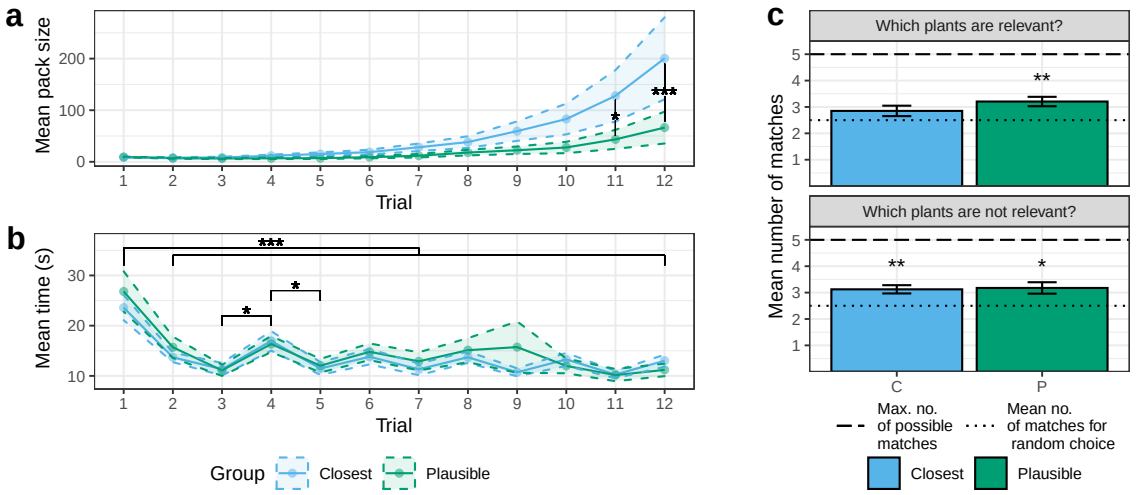

Fig. 3. Development of (**a**) mean pack size per group by trial, (**b**) mean decision time per group by trial, and (**c**) mean number of matches between user judgments and ground truth for survey items assessing relevant plants and irrelevant plants, respectively. Shaded areas in (**a**) and (**b**), and error bars in (**c**) denote the standard error of the mean. Asterisks denote statistical significance ($p < .05$ (*), $p < .01$ (**), and $p < .001$ (***), respectively. Asterisks in (**c**) denote statistical significance from expected value for random behavior.

random for both items (relevant plants: $W = 481$, $p = .005$, $r = .550$; irrelevant plants: $W = 459.5$, $p = .020$, $r = .483$). Users in the *closest* group have significantly more matches than random when identifying irrelevant ($W = 659.5$, $p = .002$, $r = .544$), but not relevant ($W = 536.5$, $p = .331$, $r = .274$) plants.

Thus, we cannot verify our hypothesis that *computationally plausible CFEs* facilitate learning. On the contrary, the development of pack size between the groups points to the opposite effect of *closest CFEs* being more beneficial for users than *computationally plausible* ones.

6.2 Do Computationally Plausible CFEs Increase User's Subjective Understanding?

To assess hypothesis 2, we analyze participant judgments on relevant survey items. Visual assessment suggests that there is very little variation in terms of user responses between groups (Figure 4a), confirmed by our statistical assessment. Groups do not statistically differ when judging whether presented CFE feedback was helpful to increase pack size (*closest* condition: $M = 3.700 \pm 1.285$ *SE*; *plausible* condition: $M = 3.636 \pm 0.242$ *SE*; $U = 656$, $p = .968$, $r = .005$). Likewise, we do not detect significant group differences in terms of subjective usability (*closest* condition: $M = 3.775 \pm 0.216$ *SE*; *plausible* condition: $M = 3.606 \pm 0.230$ *SE*; $U = 603$, $p = .513$, $r = .077$). In addition, there is no significant difference between groups for estimated usefulness of explanations for others (*closest* condition: $M = 3.750 \pm 0.208$ *SE*; *plausible* condition: $M = 3.647 \pm 0.206$ *SE*; $U = 637$, $p = .631$, $r = .056$).

6.3 Does Mode of Presentation have an Impact?

As postulated in hypothesis 3, we do not observe group differences between conditions in terms of understanding the explanations as such (Figure 4b). A considerable proportion of both groups responds positively about understanding the feedback, not differing significantly in their responses (*closest* condition: $M = 3.975 \pm 0.184$ *SE*; *plausible* condition:

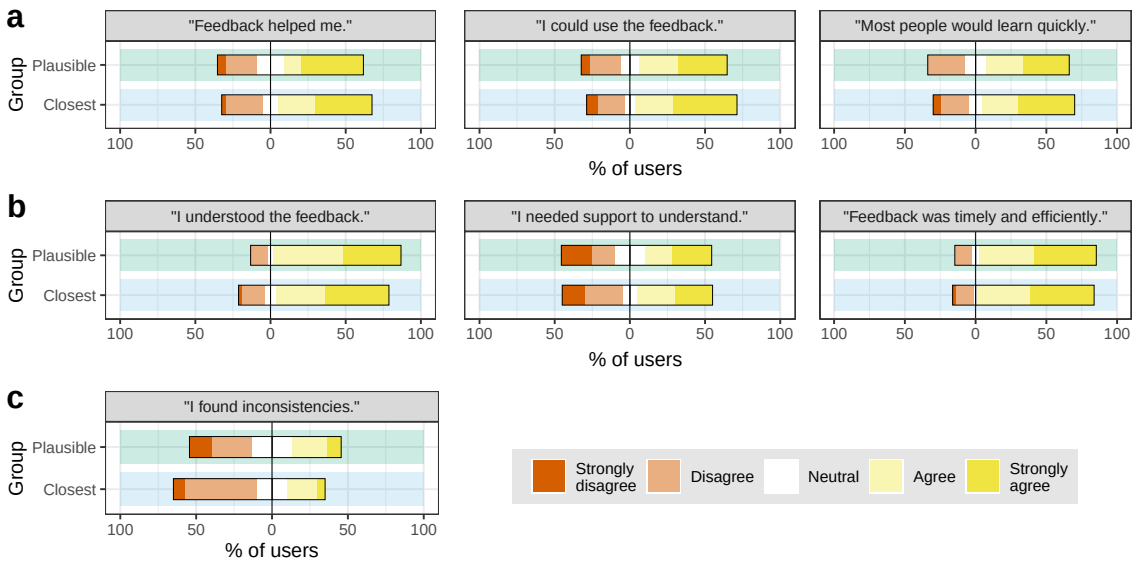

Fig. 4. Overview of user judgments in post-game survey per group, adapted from [33]. (**a**) depicts user replies in survey items relevant for hypothesis 2, (**b**) depicts user replies in survey items relevant for hypothesis 3, and (**c**) depcits replies relevant for our last exploratory analysis. Distributions did not differ significantly between groups for any of the items (all p >.05).

M = 4.118 ± 0.162 SE; U = 773.5, p = .200 r = .149). In terms of needing support for understanding, both groups reply with a similar response pattern (*closest* condition: M = 3.200 ± 0.230 SE; *plausible* condition: M = 3.147 ± 0.257 SE; U = 667, p = .890 r = .016). Similarly, user judgments on timing and efficacy of CFE presentation are consistently high across groups (*closest* condition: M = 4.100±0.175 SE; *plausible* condition: M = 4.147±0.170 SE; U = 680.5, p = 1 r = .000).

6.4 Exploratory Analysis

Our explanatory analysis revealed that more than half of all users in both groups do not report to having detected any inconsistencies in the CFEs provided (*closest* condition: M = 2.675 ± 0.166 SE; *plausible* condition: M = 2.853 ± 0.207 SE; U = 743, p = .480, r = .082).

7 DISCUSSION

In this work, we investigate effects of implementing a plausibility constraint on computed CFEs for ML models on user performance in an iterative learning task in an abstract domain. The employed constraint limits the set of possible solutions to the training data. We measure understanding and usability of explanations in terms of two objective behavioral variables, i.e., task performance and decision time, and several subjective self-reports. Our results reveal a range of valuable insights with important implications for XAI in application.

First, we cannot verify our initial supposition that *computationally plausible* CFEs facilitate learning in the current setting. Intriguingly, we observe the opposite effect: users in the *closest* condition generated larger pack sizes than users in the *computationally plausible* condition. A likely reason for this observation may be that the current study revolves around an abstract scenario of feeding aliens. Given psychological interpretations of plausibility as probability [18, 55], simply

restricting CFEs to items from the training set cannot help participants that lack an informed mental representation of the current state of this alien world. For a novice user at the onset of the study, any counterfactual is equally likely.

We may turn to the definition of psychological plausibility as comparative similarity [41, 64] for a possible explanation why users in the *closest* group showed significantly superior performance. Classically, CFEs are penalized if they deviate from the requested prediction [68], resulting in *closest CFEs* that differ minimally from the user's input. This concept resembles the view that psychologically plausible counterfactuals come from worlds that are minimally different from reality [41]. Empirical evidence highlighting the close relation between perceived plausibility and perceived similarity between counterfactual and factual state supports this notion [19, 64]. In contrast, the computational plausibility constraint rejects CFEs that are not part of the training set, even if they are minimal. Consequently, users in the *plausible* condition encounter larger differences between provided explanations and their input, at odds with the idea of plausibility as comparative similarity. Conversely, users in the closest condition might have experienced their CFEs as more psychologically plausible than the computationally plausible version. Further, an emotional effect is conceivable: Upon seeing *closest CFEs*, users might get a feeling of "just missing the better option", inducing negative affect that strongly motivates the need for improvement [47, 48]. Additionally, we may speculate that the larger discrepancies between factual and counterfactual state in the *plausible* condition may increase the mental load on users, hampering learning. Future studies need to disentangle contributions of these potential factors.

Intriguingly, our results are at odds with empirical findings indicating that CFEs for intelligent systems do not improve user's task performance [42, 66]. Lim et al. assessed the effectiveness of different explanation modes for context-aware systems [42]. In their study, performance of users receiving counterfactual style *what-if* -explanations was indistinguishable from that of users getting no explanations what-so-ever. In contrast, users in our study indeed show learning after receiving CFEs. Interestingly, their task resembles ours in so far that they also employed an abstract domain: users chose values of non-specific features (labelled A, B and C), relating to a non-specific prediction (a or b). However, while also dealing with an abstract task (i.e., feeding aliens), our users have a tangible goal (i.e., make the pack grow). Further, we refrain from separating learning and testing as in Lim et al., where users went through an initial evaluation section receiving explanation after explanation. Our design is far more interactive, with different rounds of user action and feedback. This is in line with evidence from educational science, suggesting that learner's level of engagement relates to learning outcome, with interactive activities granting deepest understanding [15]. Thus, including goal-directed and interactive settings may potentially be vital facets of effective usability studies. We suggest that future research designs need to pay special attention to these aspects in order to accurately evaluate XAI approaches.

Beyond task performance, we quantify learning success in terms of user's decision time and their ability to explicitly state which plants were crucial for their pack to prosper. Both measures do not reveal significant group differences. In terms of user's decision time, both groups show significant speed-up already after Trial 1 (Figure 3a). This initial time decrease likely reflects how participants learn to work with the game interface efficiently. Increased reaction time as a marker of learning is a classical insight from experimental psychology [44], indicating that both groups did indeed learn in the current setting. It is possible that the complex task we devised with its elaborate game-like setting was not sensitive enough or too short to pick up in subtle group differences usually linked to more simple, extensive reaction time experiments.

Users in the closest group show superior performance, however, they are not able to state more explicitly which plants were relevant or irrelevant for the given task. With this, our study replicates a recent observation that objective measures (i.e., task performance) do not necessarily correlate with self-reports reflecting system understanding [66]. Participants in both groups made three out of five correct choices on average (Figure 3c), in part significantly exceeding

the number expected in case of random behavior. Thus, both groups showed some–yet imperfect–explicit understanding of the underlying system. Potentially, users may rely on their initial mental model of the appropriate alien diet, allowing them to make advantageous feeding choices relatively quickly. However, in this initial stage, it may still be insufficient to allow clear and explicit differentiation between relevant and irrelevant features at the end of the study.

Besides effects on task performance, we do not detect any statistically meaningful differences between the two groups under investigation, predominantly affecting the evaluation of user judgments in the survey. It is clear that these observations have to be taken with the care generally devoted to null effects, calling for cautious interpretation. Still, regarding the general trends for individual survey items is informative. We cannot verify our second hypothesis, as users did not differ depending on group in terms of subjective helpfulness and usability. Still, we note that the majority of users in both groups respond with agreement or strong agreement in the respective items (Figure 4a). This supports the notion that CFEs are indeed subjectively intuitive and usable for lay users, also when used in an abstract setting.

A major challenge for effective user designs comparing different approaches is keeping conditions highly comparable, with the sole exception of the experimental manipulation. User judgments of general understandability and presentation mode of CFEs inspire confidence that we achieved this level of control with our Alien Zoo design. In fact, the respective items elicit the highest user judgments out of all survey responses, with agreement values close to 90%. High agreement across both groups leads us to conclude that mode of CFE presentation does not have an impact when comparing users experience *closest* vs. *plausible* CFEs, validating our hypothesis 3. Thus, the Alien Zoo design as proposed here is suitable for user evaluations of CFE methods, a yet vastly understudied aspect in the field of XAI [37].

Finally, the majority of users indicate that they fail to find inconsistencies in the CFEs provided. Thus, we can rule out that our code generated irregular or even contradictory explanations, confounding the observed group differences.

7.1 Limitations & Future Work

Several limitations warrant caution when interpreting our results beyond the scope of this work. Critical design choices in any XAI evaluation include the reason for explaining, and the target group [1]. The current results may only be generalized to cases with the same motivation for explaining (i.e., to 'explore') as well as the intended audience (i.e., novice users). Other motives and applications addressing more specific target groups call for independent studies.

From the originally recruited 100 participants, we excluded data from 26 individuals that did not meet our a priori data quality criteria. Such participant attrition common, especially in web-based studies. As smaller sample sizes always mean a loss of statistical power, we factored in this issue early on, based on our a priori power analysis. Yet, the effect size of the significant interaction between factors *trial number* and *group* remains relatively small. Hence, the results from this work await confirmation in larger follow-up studies.

None of the survey items revealed significant group effects, in line with a previous account of diverging trends between objective measures and self-reports [66]. This may reflect a more general tendency in human evaluation of system understanding. Alternatively, however, we may call into question the efficacy of instruments applied to assess user experience. To date, there is no standard inventory for assessing subjective usability in XAI research. We adapted the System Causability Scale [33] to determine subjective usability of presented CFEs. Yet, there is no large scale validation of this measure. One potential shortcoming may be a lack of sensitivity to subtle group differences.

Moreover, the current scenario in its present condition may be difficult to translate to specific real-world applications. The lack of realism offers full algorithmic recourse [36]: all changes are feasible (i.e., doable for the participant), and all changes in features are independent (i.e., a user can change plant 1, and this will have no long-time effect on plant 2). In real life scenarios, this is barely the case (e.g., a bank customer might never be able to get younger to get a loan; yearly

income also affects savings). Thus, our example is much more artificial, and we suggest applying iterative learning designs in more realistic, real-world scenarios as an exciting avenue for future work.

Users in our study play an elaborate online game, with a detailed user interface, and several consecutive scenes. Designed to be maximally engaging as to ensure participant compliance, it may be the case that the amount of information displayed may overwhelm some participants. This could explain inferior performance of a small proportion of participants, like those who disagree with the notion that feedback presentation was timely and efficiently.

Recent evidence suggests an added benefit of providing users with CFEs over no explanations to understand the behavior of an unknown system [66]. The current work expands this insight by a direct comparison of two different approaches for CFE computation. While our results suggests the suitability of our Alien Zoo design, further validation studies must delineate potential shortcomings. For instance, a crucial validation step of the design itself concerns comparisons of valid CFEs with no explanations or non-sensical ones.

Beyond such a fundamental investigation, the Alien Zoo design lends itself to be easily modified. Possible adaptations may incorporate data with different dynamics, use other ML models, or compare other CFE approaches. Thus, the design has tremendous potential to answer open questions in the domain of XAI. For instance, future work may explore the impact of distinct psychometric properties on performance. A small-scale user study suggests an effect of individual personality traits on user's ability to make sense of an ML system's output, and understanding the generation process, respectively [27]. It remains to be shown how personal attributes relate to usability judgments of CFEs.

Further, it is conceivable that users may prefer to receive explanations on demand, rather than continuously at prescribed intervals. There is abundant room for further progress in determining whether explicitly requesting CFEs may improve task performance, and how users would make use of their control over the explanation intervals.

Finally, we successfully demonstrate usefulness of CFEs for the current task, indicating a certain degree of intuitiveness or plausibility connected to them. Future investigations may tackle whether CFEs cause users to fall prey to a plausibility fallacy, coming to trust biased or unfair ML models just because they are coupled with intuitive explanations.

7.2 Conclusions

In this work, we present a controlled study comparing user performance and usability judgments of CFEs in an iterative learning design. We focus on potential group effects driven by receiving either *closest CFEs* that are minimally different from the user's input, compared to *computationally plausible* ones, limited to instances found in the training data. We develop an accessible game-like experimental design revolving around an abstract scenario, suitable for novice users. Our design demonstrates learning in both groups, highlighting the power of interactive and goal-directed tasks for user evaluations of CFE methods, a yet vastly understudied aspect in the field of XAI. Moreover, our findings suggest that novice users benefit more from receiving *closest* than *computationally plausible CFEs*. This supports the view of plausibility as comparative similarity rather than probability, potentially more beneficial for users if they lack an accurate mental model to build on. In sum, our work emphasizes yet again that theoretical approaches proposing explanation techniques for ML models and user-based validations thereof need to go hand in hand. Researchers designing XAI approaches need to bear in mind human behavior, preferences and mental models, to build on a solid foundation to effectively benefit the end user.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] André Artelt. 2019. CEML: Counterfactuals for Explaining Machine Learning models - A Python toolbox. https://www.github.com/andreArtelt/ceml Publication Title: GitHub repository.

[3] André Artelt and Barbara Hammer. 2019. On the computation of counterfactual explanations - A survey. *CoRR* abs/1911.07749 (2019). http://arxiv.org/abs/1911.07749 _eprint: 1911.07749.

[4] André Artelt and Barbara Hammer. 2020. Convex Density Constraints for Computing Plausible Counterfactual Explanations. In *Artificial Neural Networks and Machine Learning – ICANN 2020*, Igor Farkaš, Paolo Masulli, and Stefan Wermter (Eds.). Vol. 12396. Springer International Publishing, Cham, 353–365. https://doi.org/10.1007/978-3-030-61609-0_28

[5] André Artelt and Barbara Hammer. 2022. Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers. *Neurocomputing* 470 (Jan. 2022), 304–317. https://doi.org/10.1016/j.neucom.2021.04.129

[6] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. 2021. Evaluating Robustness of Counterfactual Explanations. *arXiv:2103.02354 [cs]* (July 2021). http://arxiv.org/abs/2103.02354

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[8] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* 67, 1 (2015). https://doi.org/10.18637/jss.v067.i01

[9] Mattan Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. 2020. effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software* 5, 56 (Dec. 2020), 2815. https://doi.org/10.21105/joss.02815

[10] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification And Regression Trees* (1 ed.). Routledge, London, UK. https://doi.org/10.1201/9781315139470

[11] Ruth M.J. Byrne. 2002. Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences* 6, 10 (Oct. 2002), 426–431. https://doi.org/10.1016/S1364-6613(02)01974-5

[12] Ruth M.J. Byrne. 2016. Counterfactual Thought. *Annual Review of Psychology* 67, 1 (Jan. 2016), 135–157. https://doi.org/10.1146/annurev-psych-122414-033249

[13] Ruth M. J. Byrne. 2007. Précis of *The* Rational Imagination: How People Create Alternatives to Reality. *Behavioral and Brain Sciences* 30, 5-6 (Dec. 2007), 439–453. https://doi.org/10.1017/S0140525X07002579

[14] Ruth M. J. Byrne, Susana Segura, Ronan Culhane, Alessandra Tasso, and Pablo Berrocal. 2000. The temporality effect in counterfactual thinking about what might have been. *Memory & Cognition* 28, 2 (March 2000), 264–281. https://doi.org/10.3758/BF03213805

[15] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (Oct. 2014), 219–243. https://doi.org/10.1080/00461520.2014.965823

[16] Louise Connell and Mark T. Keane. 2006. A Model of Plausibility. *Cognitive Science* 30, 1 (Jan. 2006), 95–120. https://doi.org/10.1207/s15516709cog0000_53

[17] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. In *International Conference on Parallel Problem Solving from Nature*, Vol. 12269. Springer International Publishing, Cham, 448–469. https://doi.org/10.1007/978-3-030-58112-1_31

[18] F. De Brigard, D.R. Addis, J.H. Ford, D.L. Schacter, and K.S. Giovanello. 2013. Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia* 51, 12 (Oct. 2013), 2401–2414. https://doi.org/10.1016/j.neuropsychologia.2013.01.015

[19] Felipe De Brigard, Paul Henne, and Matthew L. Stanley. 2021. Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility. *Cognition* 209 (April 2021), 104574. https://doi.org/10.1016/j.cognition.2020.104574

[20] Felipe De Brigard, Karl K. Szpunar, and Daniel L. Schacter. 2013. Coming to Grips With the Past: Effect of Repeated Simulation on the Perceived Plausibility of Episodic Counterfactual Thoughts. *Psychological Science* 24, 7 (July 2013), 1329–1334. https://doi.org/10.1177/0956797612468163

[21] Raphael Mazzine Barbosa de Oliveira and David Martens. 2021. A Framework and Benchmarking Study for Counterfactual Generating Methods on Tabular Data. *Applied Sciences* 11, 16 (Aug. 2021), 7274. https://doi.org/10.3390/app11167274

[22] Michelle A. Detry and Yan Ma. 2016. Analyzing Repeated Measurements Using Mixed Models. *JAMA* 315, 4 (Jan. 2016), 407. https://doi.org/10.1001/jama.2015.19394

[23] James E. Dixon and Ruth M. J. Byrne. 2011. "If only" counterfactual thoughts about exceptional actions. *Memory & Cognition* 39, 7 (Oct. 2011), 1317–1331. https://doi.org/10.3758/s13421-011-0101-4

[24] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608* (Feb. 2017). http://arxiv.org/abs/1702.08608

[25] Kai Epstude and Neal J. Roese. 2008. The Functional Theory of Counterfactual Thinking. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 12, 2 (May 2008), 168–192. https://doi.org/10.1177/1088868308316091

[26] Vittorio Girotto, Paolo Legrenzi, and Antonio Rizzo. 1991. Event controllability in counterfactual thinking. *Acta Psychologica* 78, 1-3 (Dec. 1991), 111–133. https://doi.org/10.1016/0001-6918(91)90007-M

[27] Lydia P. Gleaves, Reva Schwartz, and David A. Broniatowski. 2020. The Role of Individual User Differences in Interpretable and Explainable Machine Learning Systems. *arXiv:2009.06675 [cs]* (Sept. 2020). http://arxiv.org/abs/2009.06675

[28] Stephen D. Goldinger, Heather M. Kleider, Tamiko Azuma, and Denise R. Beike. 2003. "Blaming The Victim" Under Memory Load. *Psychological Science* 14, 1 (Jan. 2003), 81–85. https://doi.org/10.1111/1467-9280.01423

[29] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *arXiv:1805.10820 [cs]* (May 2018). http://arxiv.org/abs/1805.10820

[30] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (Jan. 2019), 1–42. https://doi.org/10.1145/3236009

[31] Fritz Heider. 1958. *The psychology of interpersonal relations.* John Wiley & Sons Ltd., New York, NY, US.

[32] Denis J. Hilton and Ben R. Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review* 93, 1 (1986), 75–88. https://doi.org/10.1037/0033-295X.93.1.75

[33] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *KI - Künstliche Intelligenz* 34, 2 (Jan. 2020), 193–198. https://doi.org/10.1007/s13218-020-00636-z

[34] P. N. Johnson-Laird and Ruth M. J. Byrne. 2002. Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review* 109, 4 (2002), 646–678. https://doi.org/10.1037/0033-295X.109.4.646

[35] Daniel Kahneman and Amos Tversky. 1982. The simulation heuristic. In *Judgment under Uncertainty* (1 ed.), Daniel Kahneman, Paul Slovic, and Amos Tversky (Eds.). Cambridge University Press, Cambridge, UK, 201–208. https://doi.org/10.1017/CBO9780511809477.015

[36] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050* (2020).

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

[38] Eugenia Kulakova, Markus Aichhorn, Matthias Schurz, Martin Kronbichler, and Josef Perner. 2013. Processing counterfactual and hypothetical conditionals: An fMRI investigation. *NeuroImage* 72 (May 2013), 265–271. https://doi.org/10.1016/j.neuroimage.2013.01.060

[39] Leah Kumle, Melissa L.-H. Võ, and Dejan Draschkow. 2021. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods* 53, 6 (May 2021), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

[40] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2019. Issues with post-hoc counterfactual explanations: a discussion. *CoRR* abs/1906.04774 (2019). http://arxiv.org/abs/1906.04774 _eprint: 1906.04774.

[41] David Lewis. 1973. Counterfactuals and Comparative Possibility. In *IFS*, William L. Harper, Robert Stalnaker, and Glenn Pearce (Eds.). Springer Netherlands, Dordrecht, 57–85. https://doi.org/10.1007/978-94-009-9117-0_3

[42] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why* and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Boston MA USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[43] Peter Lipton. 1990. Contrastive Explanation. *Royal Institute of Philosophy Supplement* 27 (March 1990), 247–266. https://doi.org/10.1017/S1358246100005130

[44] Gordon D Logan. 1992. Shapes of Reaction-Time Distributions and Shapes of Learning Curves: A Test of the Instance Theory of Automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18, 5 (1992), 883–914.

[45] Tania Lombrozo. 2012. Explanation and Abductive Inference. In *The Oxford Handbook of Thinking and Reasoning*, Keith J. Holyoak and Robert G. Morrison (Eds.). Oxford University Press, Oxford, UK, 260–276. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014

[46] Arnaud Van Looveren and Janis Klaise. 2019. Interpretable Counterfactual Explanations Guided by Prototypes. *CoRR* abs/1907.02584 (2019). http://arxiv.org/abs/1907.02584 _eprint: 1907.02584.

[47] Keith D. Markman and Matthew N. McMullen. 2003. A Reflection and Evaluation Model of Comparative Thinking. *Personality and Social Psychology Review* 7, 3 (Aug. 2003), 244–267. https://doi.org/10.1207/S15327957PSPR0703_04

[48] Victoria Husted Medvec and Kenneth Savitsky. 1997. When doing better means feeling worse: The effects of categorical cutoff points on counterfactual thinking and satisfaction. *Journal of Personality and Social Psychology* 72, 6 (1997), 1284–1296. https://doi.org/10.1037/0022-3514.72.6.1284

[49] Dale T. Miller and Saku Gunasegaram. 1990. Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology* 59, 6 (1990), 1111–1118. https://doi.org/10.1037/0022-3514.59.6.1111

[50] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[51] Chelsea Muth, Karen L. Bales, Katie Hinde, Nicole Maninger, Sally P. Mendoza, and Emilio Ferrer. 2016. Alternative Models for Small Samples in Psychological Research: Applying Linear Mixed Effects Models and Generalized Estimating Equations to Repeated Measures Data. *Educational and Psychological Measurement* 76, 1 (Feb. 2016), 64–87. https://doi.org/10.1177/0013164415580432

[52] Fabian Offert. 2017. "I know it when I see it". Visualization and Intuitive Interpretability. *arXiv:1711.08042 [stat]* (Dec. 2017). http://arxiv.org/abs/1711.08042

[53] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks against Machine Learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, Abu Dhabi United Arab Emirates, 506–519. https://doi.org/10.1145/3052973.3053009

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[55] Kathy Pezdek, Iris Blandon-Gitlin, Shirley Lam, Rhiannon Ellis Hart, and Jonathan W. Schooler. 2006. Is knowing believing? The role of event plausibility and background knowledge in planting false beliefs about the personal past. *Memory & Cognition* 34, 8 (Dec. 2006), 1628–1635. https://doi.org/10.3758/BF03195925

[56] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodriguez, Tijl De Bie, and Peter A. Flach. 2019. FACE: Feasible and Actionable Counterfactual Explanations. *CoRR* abs/1909.09369 (2019). http://arxiv.org/abs/1909.09369 _eprint: 1909.09369.

[57] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[58] Neal J. Roese. 1997. Counterfactual thinking. *Psychological Bulletin* 121, 1 (1997), 133–148. https://doi.org/10.1037/0033-2909.121.1.133

[59] Neal J. Roese and Kai Epstude. 2017. The Functional Theory of Counterfactual Thinking: New Evidence, New Challenges, New Insights. In *Advances in Experimental Social Psychology*. Vol. 56. Elsevier, Amsterdam, Netherlands, 1–79. https://doi.org/10.1016/bs.aesp.2017.02.001

[60] Lawrence J. Sanna and Kandi Jo Turley. 1996. Antecedents to Spontaneous Counterfactual Thinking: Effects of Expectancy Violation and Outcome Valence. *Personality and Social Psychology Bulletin* 22, 9 (Sept. 1996), 906–919. https://doi.org/10.1177/0146167296229005

[61] Maximilian Schleich, Zixuan Geng, Yihong Zhang, and Dan Suciu. 2021. GeCo: Quality Counterfactual Explanations in Real Time. *arXiv:2101.01292 [cs]* (May 2021). http://arxiv.org/abs/2101.01292

[62] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, New York, NY, USA.

[63] Barry Smyth and Mark T. Keane. 2021. A Few Good Counterfactuals: Generating Interpretable, Plausible and Diverse Counterfactual Explanations. *arXiv:2101.09056 [cs]* (Jan. 2021). http://arxiv.org/abs/2101.09056 arXiv: 2101.09056.

[64] Matthew L. Stanley, Gregory W. Stewart, and Felipe De Brigard. 2017. Counterfactual Plausibility and Comparative Similarity. *Cognitive Science* 41 (May 2017), 1216–1228. https://doi.org/10.1111/cogs.12451

[65] Ilia Stepin, Alejandro Catala, Martin Pereira-Fariña, and Jose M. Alonso. 2019. Paving the way towards counterfactual generation in argumentative conversational agents. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. Association for Computational Linguistics, Tokyo, Japan, 20–25. https://doi.org/10.18653/v1/W19-8405

[66] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (Feb. 2021), 103404. https://doi.org/10.1016/j.artint.2020.103404

[67] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual Explanations for Machine Learning: A Review. *arXiv:2010.10596 [cs, stat]* (Oct. 2020). http://arxiv.org/abs/2010.10596

[68] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841. Publisher: HeinOnline.

[69] C.R. Walsh and Ruth M.J. Byrne. 2005. The mental representation of what might have been. In *The psychology of counterfactual thinking*, D.R. Mandel, Denis J. Hilton, and P. Catellani (Eds.). Routledge, London, 61–73.

[70] Adam White and Artur d'Avila Garcez. 2020. Measurable Counterfactual Local Explanations for Any Classifier. In *ECAI*. Santiago de Compostela, Spain, 7.