The plausible Alien Zoo: Users prefer closest, and not plausible counterfactual explanations

Subtitle

ANONYMOUS AUTHOR(S)

This will be the abstract at some point.

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

ACM Reference Format:

Anonymous Author(s). 2021. The plausible Alien Zoo: Users prefer closest, and not plausible counterfactual explanations: Subtitle. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2022).* ACM, New York, NY, USA, 12 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 [15]. 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 advent of explainable artificial intelligence (XAI) as a novel research field. Consequently, Consequently, the XAI community has seen a veritable surge of technical accounts on how to realize explainability for ML [14].

Motivated by a seminal review by [23] advocating a user-centered focus on explainability, counterfactual explanations (CFEs) gained particular prominence as a supposedly useful, human-accessible solution [18]. 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 explainability of ML models more generally, shows an alarming tendency to take the quality of the suggested explanation modes at face value [10, 25]. In a recent review of over 100 counterfactual XAI studies, Keane et al. criticize that only a third of these studies concern themselves with user-based evaluations, often with great limitations concerning statistical power and reproducibility [18].

The lack of user-based evaluations affects not only assessments CFEs per se, 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 [4], 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.

Manuscript submitted to ACM

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 [3, 30, 31], no user study to date has directly investigated potential benefits of enforcing an additional plausibility constraint on generating CFEs. Therefore, we perform a well-powered user study analyzing the performance of novice users when receiving conventional CFEs exclusively defined via their proximity to the decision boundary, compared to algorithmically plausible CFEs as feedback in an iterative learning design [2, 3].

The remainder of this paper is structured as follows: We will first briefly outline CFEs as a psychologically grounded mode of explanation, with special regard to the impact of plausibility in terms of counterfactual thought (Section 2). Section 3 gives a more technical account of conventional—henceforth referred to as close—and plausible model agnostic CFE approaches. After delineating guiding questions and hypotheses for our research (Section 4), we give an in-depth account of our suggested experimental design, as well as the experimental procedure of our user study (Section 5). Results are depicted in Section 6, before closing with an in-depth discussion of insights drawn from this study, including limitations and avenues for future work in Section 7.

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 explanation may depend on the reason why we need to explain in the first place [1], as well as pre-existing knowledge and experiences of users receiving a certain explanation [33]. 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 [23]. A central insight from classical psychological literature is that human explanations are typically contrastive: They emphasize (explicitly or implicitly) why a specific outcome occured instead of another [16, 20, 22, 23]. This contrastive nature is intrinsically linked to the more general human tendency to reflect upon past events by generating possible alternatives, i.e., counterfactual thinking [27]. Empirical evidence suggests that humans show this *what-if* mentality spontaneously [12], and increasingly when facing negative outcomes or unexpected results [29]. In their *Functional Theory of Counterfactural Thinking*, Roese and Epstude suggest a crucial role of counterfactual thought to guide to formation of future intentions, thus regulating subsequent behavior [11, 28]. 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 [2, 3, 8, 13, 32].

2.1 The advantage of plausibility / Humans generate plausible CFEs

What is plausibility? PEZDEK, 2006, Memory & Cognition: seem to equate plausibility with probability Definitions of plausibility have ranged from the abstract and immeasurable, ‘having intuitive logic’, to the narrow and measurable, “how far we go into the tails of the distribution” (Breuer et al., 2009). Breuer, T., Jandacka, M., Rheinberger, K. and Summerb, M. (2009) ‘How to find plausible, severe, and useful stress scenarios’, International Journal of Central Banking, productive relationship between plausibility and probability of scenarios: Millett, 2009 -Millett, S.M. (2009) ‘Should probabilities be used with scenarios?’, Journal of Futures Studies Morgan and Keith: (2008, p.196; Climatic Change), for example, assert: “The literature on scenarios often aims to make a sharp distinction between scenarios and forecasts or projections; for example, it is asserted that scenarios are judged by their ‘feasibility’ or ‘plausibility’ rather than their likelihood. We cannot find any sensible interpretation of these terms other than as synonyms for relative subjective probability. Absent a supernatural ability to foresee the future, what could be meant by a statement that one scenario is feasible and another infeasible but that the first is (subjectively) more probable than the second?” "plausibility seeks to prepare for a variety of future states that are considered ‘occurrable’ (could happen), explicitly including some that are not the most likely ones." Wiek, 2013, Plausibility indications in future scenarios from keane, 2021: "We also set aside the fact that the concept of plausibility is not well understood in Psychology; though, there is agreement that it tends to depend heavily on user’s knowledge of a domain, rather than on similarity per se [Connell and Keane, 2007]." "Byrne 2016: plausible—reasonable, believable, and acceptable." – "People create counterfactuals that are plausible. They exhibit remarkable regularities in the alternatives to reality that they create. Most people zoom in on the same pivotal junc- tures or fault lines in their representation of reality to create a counterfactual alternative to it." Inutitively, humans tend to focus on changing antecedents that are probable: people do not tend to create fantastical CFs, but plausible ones (CHECK: De Brigard et al., 2013) also: controllable: modify what is controllable [Girotto et al., 2007] (note however: exceptions [Girotto et al., 2007], [Pighin et al., 2011]) also: plausibility is closely connected to comparative similarity - "counterfactuals considered to be true in possible worlds comparatively more similar to ours are judged as more plausible than counterfactuals deemed true in possible worlds comparatively less similar." - based on David Lewis (1973, 1979) work; formulated by Stanley, DOI: 10.1111/cogs.12451, 2017 - "Our results suggest that the greater the perceived similarity between the original memory and the episodic counterfactual event, the greater the perceived plausibility that the counterfactual event might have occurred." Same group, similar insight: De Brigard plausibility may be dangerous: one may assume that something is true, just because it is plausible (i.e., the plausibility fallacy) If plausibility would be incorporated in the CFEs, it could help.

3 COMPUTATION OF CFS AND PLAUSIBLE CFS

4 DO USERS PROFIT FROM ALGORITHMIC PLAUSIBILITY?

The main research question of the current work is: Can plausible CFEs help users? Further Guiding questions: Do plausible CFE facilitate learning? Do plausible CFEs increase user’s subjective understanding? Given the robustness advantage of plausible CFEs over closest CFEs, and given that we know that plausibility helps users understand (TO DO!), we formulated the following 4 hypotheses. First, we expect plausible CFEs to be more helpful to users tasked to discover unknown relationships in data than closest ones, both in terms of objective and subjective understandability (hypothesis 1). Specifically, we expect participants in the plausible condition a) to produce larger packs over time, b) to become more automatic and thus quicker in choosing their plants, and c) to clearly state which plants were crucial for their pack to prosper.

Second, we expect to see difference in terms of subjective understanding between both groups (hypothesis 2). We predict that users will differ in how far they find the explanations useful, and in how far they can make use of it, with an advantage of plausible CFEs. We also posit that users imagine plausible CFEs to be more helpful for others users.

An important feature of the current design is the high comparability of conditions, as structure and presentation mode of CFEs is kept constant for both groups. Thus, we do not expect users in different conditions to differ in their understanding of the explanations per se, their need for support to understand, and their evaluation of timing and efficacy of CFE presentation (hypothesis 3).

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

5 EXPERIMENTAL DESIGN

To test hypothesis1,hypothesis2,hypothesis3, ... from 4, we... used the AlienZoo paradigm (REF to arxiv when it's there).

5.1 The plausible Alien Zoo paradigm

In the current study, be relied on the novel Alien Zoo paradigm (REF to arxiv paper). In this scenario, participants are asked to imagine themselves as zoo keepers in a zoo for aliens. Several alien plants may be chosen to feed to the aliens, but it is not clear what plants make up a nutritious diet for the aliens. Thus, participants are tasked to find how to best feed the aliens. Participants go through several feeding cycles, choosing a combination of plants for feeding. 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 counterfactual explanation). (For details on the procedure, see 5.6). Advantages of this approach is control for individual domain knowledge by relying on an abstract domain, evaluating the objective as well as subjective usability of automatically generated explanations directly (REF to arxiv paper). Thus, this current use case corresponds to a human grounded evaluation following the taxonomy by Doshi-Velez and Kim, assessing perforemance of real users in an abstract task setting. Unlike most real-life applications, systems investigated in user studies typically focus on applications for expert users trying to gain expert knowledge of a domain. In contrast, we investigate whether providing CFEs to novice users improves their understanding of relationships in a yet unknown dataset. Thus, our setting falls under the explaining to discover reasons for explainability defined by Adadi and Berrada [1]. Specifically, we focus on the question whether CFEs computed according to algorithmic plausibility prove to be more useful for novice users in this setting than providing closest CFEs, independent of domain. Relying on novice users in an abstract context, we mitigate any difference in domain knowledge and possible misconceptions about the task setting [33].

5.2 Post-Game survey

The final post-game questionnaire is used to collect self-report information from participants. Participants are asked several Likert-scale questions, inspired by the System Usability Scale [17], a Werkzeug designed to measure the quality of explanations elicited by an explainable ML system. Its purpose is to obtain an understanding of how users feel about using our system, including whether they trusted and understood the system and explanations.

5.3 Constructs, expected relations and measurements

Maybe here: set up and add a causal diagram of our study as in van der Waa (2021)? We measure understanding and usability of explanations in terms of two objective behavioral variables and several subjective self-reports. In terms of behavior, we assess the development of pack size in the Alien Zoo game over trials. As a measure of task performance, this value indicates the extend of user's understanding of relevant in irrelevant features in the underlying dataset, assuming that a solid understanding leads to better feeding choices. Second, we measure time needed to reach a feeding decision over trials. As we assume participants to become more automatic in making their plant choice, we expect this practice effect to be reflected in terms of decreasing time required to reach a feeding decision [21]. We acquired self-reported measurements using the post-game survey, testing different aspects of participant's system understanding. The first two survey items explicitly ask users to identify plants they think are relevant and irrelevant for task success. Replies from these items allow us to measure to shich extend users in different groups formed explicit knowledge of the underlying data structure. Further, users indicate how far they find the explanations useful, in how far they can make use of it, and in how far they imagine the presented CFEs to be helpful for others users, too. These items assess users subjective understanding. Finally, three self-reported measurements check for potential confounds. These are items that ask users to indicate their understanding of the explanations per se, 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 across groups, which would be a confounding variable.

5.4 ML model and implementation

The web interface was developed using [REF to phaser, javascript, etc.]. The behavior of the game is determined by predictions of underlying ML models. While technically, any ML algorithm (that is CEML cimpatible) could be used, we chose to use ...

5.5 Participants

We developed a game-like experimental design based on a web-based interface, to facilitate access for users from all over the world and diverse backgrounds, enabling large-scale and easy participant recruitment.

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 the 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 [5]. The study was approved by the Ethics Committee of the University Bielefeld, Germany.

5.6 Experimental Procedure

Here: insert image(s) of game UI and study flow.

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 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. 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 number of Shubs, the number of Shubs 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. 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 Shubs 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 plausible counterfactual. 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 number of shubs 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 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 shubs 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 [17] 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.

5.6.1 Statistical Analysis, Sample Size Calculation and Data Quality Measures. We perform all statistical analyses using R–4.1.1 [26], using CFE variant (closest or 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) [6]. In the model testing for differences in terms of user performance, the dependent variable is number of Shubs generated. In the assessment of user's reaction time, we use time needed to reach a feeding decision 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 [9, 24]. 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) [7]. 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 bonerroni corrected to account for multiple comparisons.

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)

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. We follow the same approach to compare age and gender distributions. To analyse 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 shub number and reaction time as described above. For each of these models, we run a simulation-based power analyses for samples of 20–100 participants with fixed effects of group and trial number over 1000 simulations (mixedpower v.0.1.0) [19]. 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.

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 implemented a number of data quality checks that were planned a priori. We identify speeders based on the time needed to reach a feeding decision, 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 shubs in both attention trials during the game. Furthermore, we included a catch item (item 7) in the survey, asking users to tick the "I prefer not to answer." option, flagging users that did not pay attention to the survey items. 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.

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 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, time needed to reach a feeding decision, and matches between ground truth and indicated plants. Figure 1a shows the development of average pack size as well as average time spent to reach a feeding decision 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* ($F(11{,}792) = 2.1193$, $p = 0.017$, $\eta_\text{p}^2 = 0.0286$) in the corresponding linear mixed effects model. The follow-up analysis reveals significant differences between groups in trial 11 ($t(472) = 4.040$, $p = .0.012$, $d = 0.6929$) and trial 12 ($t(472) = 2.530$, $p < .001$, $d = 1.1010$). Additionally, there is a highly significant main effect of trial number ($F(11{,}792) = 7.5851$, $p < .001$, $\eta_\text{p}^2 = 0.0953$), but no significant main effect of group ($F(11{,}72) = 2.5857$, $p = .112$, $\eta_\text{p}^2 = 0.0347$).

Participants in both groups showed a marked decrease in time needed to reach a feeding decision over the curse of the study, already apparent after the very first trial (Figure 1b). The significant main effect of factor *trial number* ($F(11{,}792) = 14.8177$, $p < .001$, $\eta_\text{p}^2 = 0.1707$) confirms this observation. Corresponding post-hoc analyses show significant differences between trial 1 and all other trials (all $t(792) > 5.90$, $p < .001$, $d > 1.2000$), between trial 3 and 4 ($t(792) = 3.765$, $p = 0.012$, $d = 0.6210$), and between trials 4 and 5 ($t(792) = 3.395$, $p = 0.048$, $d = 0.5600$). Neither the main effect of factor *group* ($F(11{,}72) = 0.2347$, $p = .630$, $\eta_\text{p}^2 = 0.0032$), nor the interaction between factors *trial number* and *group* ($F(11{,}792) = 0.8966$, $p = .543$, $\eta_\text{p}^2 = 0.0123$) reach significance.

In terms of mean number of matches between user judgements of plant relevance for task success and the ground truth, users in both groups performed comparably both for relevant (closest: mean number of matches = 2.8500 ± 0.1979 *SE*; plausible: mean number of matches = 3.2060 ± 0.1780 *SE*; $U = 781$, $p = 0.255$, $r = 0.0540$) and irrelevant plants (closest: mean number of matches = 3.125 ± 0.1568 *SE*), plausible: mean number of matches = 3.1765 ± 0.2172 *SE*; $U = 721.5$, $p = .643$, $r = 0.0539$).

Thus, we cannot verify our hypothesis that 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 plausible ones.

6.2 Do plausible CFEs increase user's subjective understanding?

To assess hypothesis 2, we analyze participant judgements on relevant survey items. Visual assessment suggests that there is very little variation in terms of user responses between groups (Figure 2a), confirmed by our statistical assessment. Groups do not statistically differ when judging whether presented CFE feedback was helpful to increase pack size (closest condition: $M = 3.7000$ ± 1.2850 *SE*; plausible condition: $M = 3.6364$ ± 0.2416 *SE*; $U = 656$, $p = .968$, $r = 0.0047$). Likewise, we do not detect significant group differences in terms of subjective usability (closest condition: $M = 3.7750$ ± 0.2163 *SE*; plausible condition: $M = 3.6061$ ± 0.2300 *SE*; $U = 603$, $p = .513$, $r = 0.0765$). In addition, there is no significant difference between groups for estimated usefulness of explanations for others (closest condition: $M = 3.7500$ ± 0.2080 *SE*; plausible condition: $M = 3.6471$ ± 0.2063 *SE*; $U = 637$, $p = .631$, $r = 0.0558$).

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 per se (Figure 2b). A considerable proportion of both groups responded positively about understanding the feedback, not differing significantly in their responses (closest condition: $M = 3.9750$ ± 0.1843 *SE*; plausible condition: $M = 4.1176$ ± 0.1622 *SE*; $U = 773.5$, $p = .200$ $r = 0.1489$). In terms of needing support for understanding, both groups reply with a similar response pattern (closest condition: $M = 3.2000$ ± 0.2298 *SE*; plausible condition: $M = 3.1471$ ± 0.2573 *SE*; $U = 667$, $p = .890$ $r = 0.0161$). Similarly, user judgements on timing and efficacy of CFE presentation are consistently

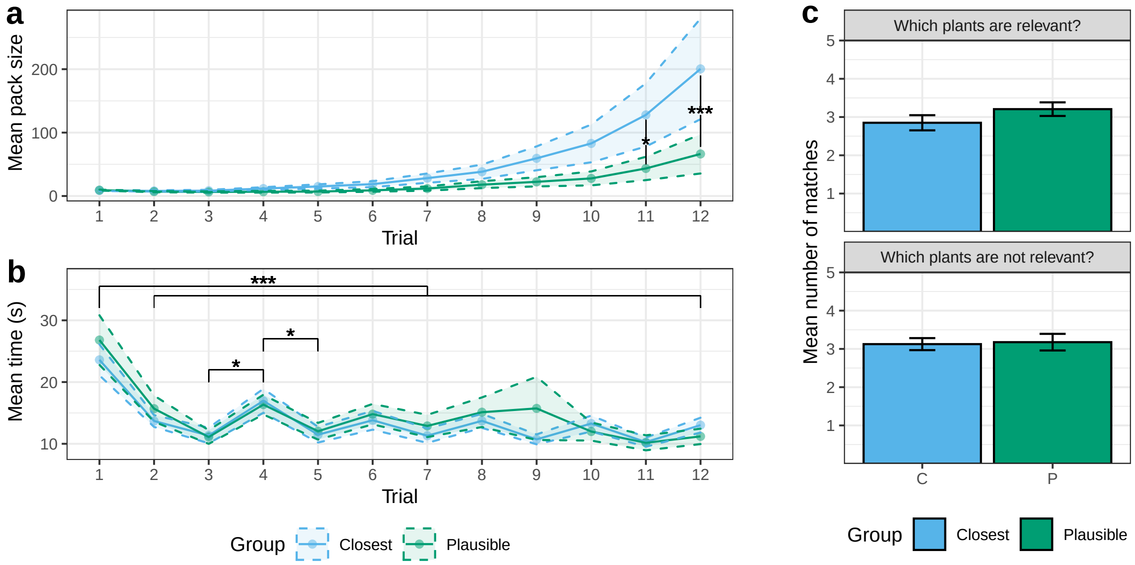

Fig. 1. Development of **a** mean pack size per group by trial, **b** mean time needed to reach a feeding decision per group by trial, and **c** mean number of matches between user judgements and ground truth for survey items assessing relevant plants and irrelevant plants, respecively. Shaded areas in **a** and **b**, and error bars in **c** denote the standard error of the mean. Asterisks denote statistical significance at $p < 0.05$ (*) and $p < .001$ (***), respectively.

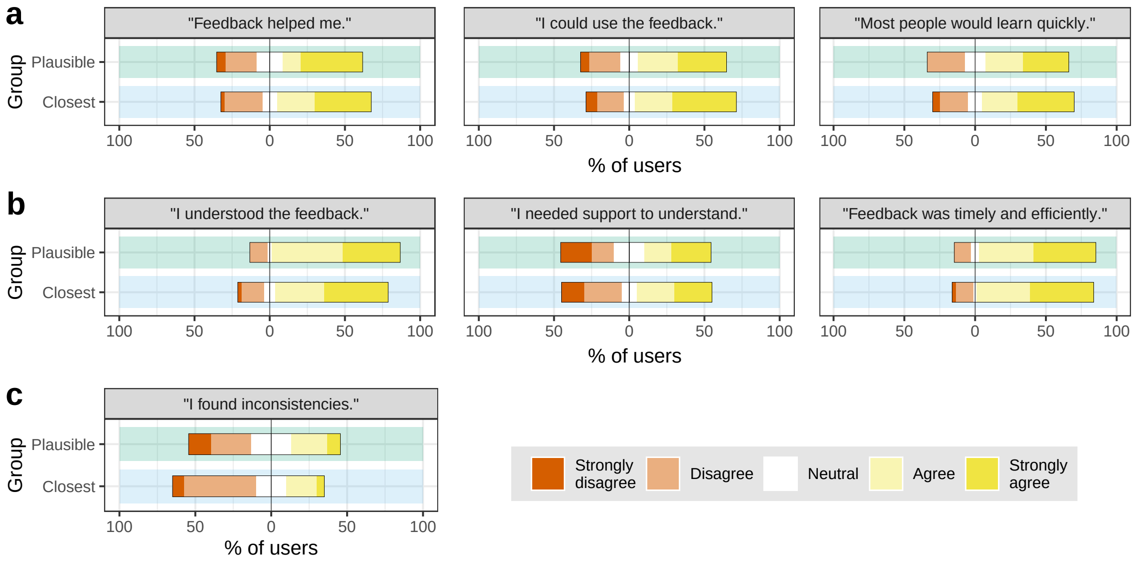

Fig. 2. Overview of user judgements in post-game survey per group, adapted from [17]. **a** depicts user replies in survey items relevant for hypothesis 2, **b** depicts user replies in survey items relevant for hypothesis 3, and **c** depcits replies relevant for our last exploratory analysis. Distributions did not differ significantly between groups for any of the items (all p >0.05).

high across groups (closest condition: M = 4.1000±0.1747 SE; plausible condition: M = 4.1471±0.1696 SE; U = 680.5, p = 1 r = 0.0000).

6.4 Exploratory Analysis

Our explanatory analysis revealed that more than half of all users in both groups did not detect inconsistencies in the CFEs provided (closest condition: M = 2.6750 ± 0.1656 SE; plausible condition: M = 2.8529 ± 0.2074 SE; U = 743, p = 0.480 r = 0.0820).

7 DISCUSSION

Here's a thing: It is notable that a considerable proportion of both groups responded positively about understanding the feedback, supporting the general notion that CFEs are very human friendly. The highest agreement is inpresentation of feedback (timely and efficiently), which is a win for our design! What we can also show: the framework of the alien Zoo is suitable to test the usability of a particular CF method. We could also use it to test different CF methods agains each other! Such comparative user studies are sparse: Akula, et al., 2020; Förster et al., 2020a, 2020b (referenced in Keane, 2021) Lim 2019 writes: "Our results may suggest the ineffectiveness of How To and What If explanations, but these intelligibility types may be more useful for other types of tasks, particularly those relating to figuring out how to execute certain system functionality, rather than interpreting or evaluating." - are we better? Why? What could lead to problems? From Lage, 2019: "Kulesza et al. (2013) performed a qualitative study in which they varied the soundness and the completeness of an explanation of a recommendation system. They found completeness was important for participants to build accurate mental models of the system." – Are CFs really complete? Is it guaranteed that user's get the complete picture given a subset of explanations highly dependent on their own input? Lage et al., 2019: CF evaluation is harder than mere simulation or verification tasks (i.e., users need significantly longer and were judged as being significantly harder to do than simple simulations) Check out: improved task performance are contingent on both system understanding and appropriate trust (CHEF REF: R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: challenges and prospects, arXiv preprint arXiv:1812.04608.) in van der Waa 2021: explanations did not improve task performance (choosing the correct dosage of insulin) - why did it (hopefully) work for us? Building upon work by Isabel Valera: in our case, plausibility corresponds to algorithmic recourse, as all changes in features are independent (user can change leaf 1, and this will have no long-time effect on leaf 2), and all changes are feasible (i.e., doable for the participant). In real life, this is not always the case (a bank customer might never be able to get younger to get a loan; yearly income also affects savings) - thus, our example is more artificial Our findings are a replication of a phenomenon observed by van der Waa, 2021: in a recent study, van der Waa showed that perceived system understandability did not correlate with any objective measure of performance, shedding profound doubt onto whether we should just assume something to be understandable.

7.1 Limitations & Future Work

Two critical design decisions need to be considered are the reason for explaining, and who the given explanation targets. Thus, it is important to keep in mind, that ganeralizability of results of user based studies is likely limited to the current reason (to 'explore') as well as the current target audience (novice users). Our experimental design caters to one need for explanation only. We can only make statements as to how CFs are useful when ML predictions are used to learn something about the system. This might prove super helpful for ML tutoring systems. However, there are more reasons to explain (debugging, etc., see Adadi again) that the current design cannot cover. Future work and other settings are needed. Out design has the advantage to be very generall, addressing the lay user. For CFs that are supposed to assist experts in very specific applications, e.g., in the medical (REF) field or so (others?) need more specific user assessments and experimental designs targeting experts. From a psychological standpoint, it is important to note that definitions of CFs slightly diverge. Many computer science definitions consider a CF as valid only if it represents a ”minimal set of features to change” (Ref?). In Psychology, however, any combination of changes leading to a different outcome can be considered a counterfactual (true, there are better and worse ones, but still...) Out study was set up to be engaging - but it is also complex; there is a chance that users could have been overwhelmed with the user interface Impact of individiual traits on performance here could not be tested (see Gleaves et al., 2020). Future work! Abstract domain - probably not suitable for other specific domain such as CFs for image classification Many XAI studies supposedly suffer from confirmation bias - does our do so, too?: Rosenfeld, A. (2021, May). Better Metrics for Evaluating Explainable Artificial Intelligence. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (pp. 45-50). Another cool design worth looking at: chatbot use! Chat-XAI: A New Chatbot to Explain Artificial Intelligence, 2021; Gao,M.,Liu,X.,Xu,A.,&Akkiraju,R.(2021,September).Chat-XAI: ANewChatbottoExplainArtificialIntelligence.InProceedingsofSAIIntelligentSystemsConference(pp.125-134).Springer,Cham. Future work: Our design has tremendous potential to be adapted to answer more questions... Lim, 2009: ”Another issue with real systems is that users may not like to receive explanations all the time, but on demand instead, because the former may be too obtrusive. We would like to run a future study to compare if users can still benefit sufficiently from explanations if they get to choose when and how often they can receive explanations, and if explicit effort in asking for explanations can improve learning.” Adapt this point for our purpose as well! Whithin-subject study would be cool! Look at various age groups (including kids!) and explore the effect other One thing in our case: "plausibility is usually confounded with degree of background knowledge." (PEZDEK, 2006, Memory & Cognition) - none of our people had background knowledge (sold as an advatage so far)

7.2 Conclusion

In this work, we present a large controlled study of... We developed a game-like experimental design based on a web-based interface... The code of this is available, making it easy to adapt to other approaches to compute CFs and other underlying ML models. We welcome its use by other research groups and practitioners, to further advance the currently limited insights into the field. Our findings suggest that providing (plausible) counterfactuals explanations to novice users...

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 ML019.

[2] André Artelt. 2020. Counterfactual explanations of ML models.

[3] 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

[4] 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 arXiv: 2103.02354.

[5] 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

[6] 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

[7] 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

[8] 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 Series Title: Lecture Notes in Computer Science.

[9] 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

[10] 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 ML001.

[11] 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

[12] 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

[13] 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 arXiv: 1805.10820.

[14] 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

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

[16] 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

[17] 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

[18] 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 arXiv: 2103.01035.

[19] 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

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

[21] 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.

[22] 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, 260–276. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014

[23] 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

[24] 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

[25] Fabian Offert. 2017. "I know it when I see it". Visualization and Intuitive Interpretability. In *arXiv:1711.08042 [stat].* Long Beach, CA, USA. http://arxiv.org/abs/1711.08042 arXiv: 1711.08042.

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

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

[28] 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, 1–79. https://doi.org/10.1016/bs.aesp.2017.02.001

[29] 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

[30] 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 arXiv: 2101.01292.

[31] 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.

[32] 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

[33] 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 ML022.