Categorical and Continuous Features in Counterfactual Explanations of AI Systems

Greta Warren
greta.warren@ucdconnect.ie
School of Computer Science, University College Dublin
Dublin, Ireland

Ruth M.J. Byrne
rmbyrne@tcd.ie
School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin
Dublin, Ireland

Mark T. Keane
mark.keane@ucd.ie
Insight SFI Centre for Data Analytics, VistaMilk SFI Research Centre, School of Computer Science, University College Dublin
Dublin, Ireland

ABSTRACT

Recently, eXplainable AI (XAI) research has focused on the use of counterfactual explanations to address interpretability, algorithmic recourse, and bias in AI system decision-making. The proponents of these algorithms claim they meet users' requirements for counterfactual explanations. For instance, many claim that the output of their algorithms work as explanations because they prioritise "plausible", "actionable" or "causally important" features in their generated counterfactuals. However, very few of these claims have been tested in controlled psychological studies, and we know very little about which aspects of counterfactual explanations help users to understand AI system decisions. Furthermore, we do not know whether counterfactual explanations are an advance on more traditional causal explanations that have a much longer history in AI (in explaining expert systems and decision trees). Accordingly, we carried out two user studies to (i) test a fundamental distinction in feature-types, between categorical and continuous features, and (ii) compare the relative effectiveness of counterfactual and causal explanations. The studies used a simulated, automated decision-making app that determined safe driving limits after drinking alcohol, based on predicted blood alcohol content, and user responses were measured objectively (users' predictive accuracy) and subjectively (users' satisfaction and trust judgments). Study 1 (N=127) showed that users understand explanations referring to categorical features more readily than those referring to continuous features. It also discovered a dissociation between objective and subjective measures: counterfactual explanations elicited higher accuracy of predictions than no-explanation control descriptions but no higher accuracy than causal explanations, yet counterfactual explanations elicited greater satisfaction and trust judgments than causal explanations. Study 2 (N=211) found that users were more accurate for categorically-transformed features compared to continuous ones, and also replicated the results of Study 1. The findings delineate important boundary conditions for current and future counterfactual explanation methods in XAI.

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03.
https://doi.org/10.1145/3581641.3584090

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **HCI theory, concepts and models**; • **Computing methodologies** → *Cognitive science.*

KEYWORDS

XAI, explanation, counterfactual, user study

ACM Reference Format:
Greta Warren, Ruth M.J. Byrne, and Mark T. Keane. 2023. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3581641.3584090

1 INTRODUCTION

The burgeoning prevalence of automated decision-making in the public and private sectors has led to increased concerns about the fairness, transparency, and trustworthiness of these Artificial Intelligence (AI) systems [3, 10]. Automated counterfactual explanations have emerged as a common strategy to help users understand the decisions of such systems and address fairness and trust issues [26, 28]. Such explanations describe how an AI system's output decision would have been different, had some of the input features been different. The typical example of a counterfactual explanation is one in which a bank customer is refused a loan by an automated system and on querying the decision is told "if you had asked for $2k less, you would have received the loan". Counterfactuals are viewed as a promising solution to the eXplainable AI (XAI) problem because of their compliance with data protection regulations (*e.g.*, EU GDPR [68]), their potential to support algorithmic recourse [27], and their acknowledged importance in human explanations [7, 54]. However, although there is now a substantial XAI literature reporting many diverse counterfactual algorithms, there is a paucity of good user studies that back the claims made for this explanation strategy [28].

We do not know precisely how people understand counterfactual explanations of AI decisions, what impact these explanations have on people's knowledge of the AI system or domain, and which aspects of counterfactual methods are psychologically critical to successful XAI. For instance, many counterfactual algorithms prioritise altering "plausible", "actionable", or "causally-important" features in their generated explanations [25, 55, 68]. However, it has not been reliably established which of these feature-types (if any) are important to users. In the present study, we examine a feature-type

distinction that is arguably fundamental, that between continuous features and categorical features. Although counterfactual methods differ in how they compute these feature-types, they have not been identified as being critically important. Yet, psychological evidence suggests that people do not tend to construct counterfactuals that modify continuous features [23].

We also do *not* know whether counterfactual explanations offer significant advantages over other long-standing explanation strategies, such as causal rules (*e.g.*, in expert systems [5] and decision trees [17, 39]). In philosophy [20, 42], it has been argued that causal and counterfactual reasoning are closely related. Furthermore, psychological evidence has shown that counterfactual thinking plays a key role in people's understanding of causality [15, 50]. However, counterfactual and causal explanations have seldom been compared to one another in XAI. Here, we explicitly compare these two explanation strategies, under carefully controlled and matched conditions, to determine their relative impacts on people's understanding and judgments.

Hence, the present studies examine the effects of categorical and continuous features in counterfactual and causal explanations for users of a simulated automated decision-making app that determines drink-driving limits, based on a model's predictions of people's blood alcohol content (BAC). Different groups of users were given the app's decisions, with either counterfactual explanations *(e.g.*, "if John had drunk 3 units, he would have been under the limit"), or causal explanations (*e.g.*, "John was over the limit because he drank 5 units"), or no-explanation, that is, they were presented with a re-description of the app's decision (*e.g.*, "John was over the limit"). Experiment 1 (N=127) examined the effects of explanations using categorical features (gender, stomach-fullness) and continuous features (units, duration of drinking, body weight) on people's accuracy in predicting the system's decisions and their subjective satisfaction and trust in the system. Experiment 2 (N=211) directly compared the impact of continuous and categorical features, for counterfactual and causal explanations, using the same measures.

In the next section, we introduce relevant related work on counterfactual and causal explanations (see 2.1), the cognitive differences between them (see 2.2), and feature-types in XAI systems (see 2.3), before outlining our experimental paradigm (see 3). We then report two experiments (see 4 and 5) and discuss the implications of our findings for different explanation strategies and feature-types and explanations in XAI (see 6).

2 RELATED WORK

In the following subsections, we review three key literatures that provide context for the current work, namely the related work on (i) counterfactual and causal explanations in XAI, (ii) cognitive differences between counterfactual and causal explanations, and (iii) consideration of different feature-types in counterfactual XAI. As we will see, much of the XAI literature emphasises the role that explanations play in establishing satisfaction and trust in an automated decision-maker. Clearly, these subjective self-reported responses are important and should be explored. However, explanations should also impact people's understanding of the system, the task, and the domain [32]. Philosophers and psychologists have repeatedly argued that true explanations must result in a change in people's understanding of the world, events or phenomena [30, 54]. In XAI, this requirement means that good explanations should improve people's understanding of the AI system, the task domain and, if required, their performance on some relevant target task [19, 32]. An explanation is effective, therefore, if people objectively perform better on a task involving the AI system, for example, by being faster, more accurate, or by being able to predict what the system might do next [16, 37, 43, 46, 65]. Another theme in the current XAI literature is the divergence between people's objective understanding of an AI system and their subjective assessments of it, that is, a lack of correspondence between objective and subjective measures, and we consider this divergence in the next sections also.

2.1 Counterfactual and Causal Explanation: Evidence from User Studies

A counterfactual explanation explains an outcome, *e.g.*, that John is over the blood-alcohol limit to drive, by describing how an alternative outcome would have occurred, had an antecedent event been different, e.g., John would have been under the limit if he had drunk 3 units of alcohol. It explains the facts, *e.g.*, that John drank 6 units and was over the limit, by contrasting them with a counterfactual alternative case with different feature-values and a different output, *e.g.*, John drank 3 units and was under the limit. In XAI, counterfactual explanations typically advise a user about how a different system output would have been achieved by some change to the input features, *e.g.*, a customer would have been granted a loan if their credit score had been higher. In contrast, a causal explanation explains an outcome by identifying the antecedent event that led to it, *e.g.*, John was over the limit because he drank 6 units. It refers only to the facts and does not propose how changes to the facts would lead to a different outcome. In this section, we examine how counterfactual and causal explanations have been used in XAI, and we review their effects on human users' understanding of AI systems.

A recent upsurge in XAI research on counterfactual explanation has produced ~125 distinct counterfactual algorithms (for reviews see [26, 28, 67]). These methods advance different algorithms for computing counterfactuals, with approaches ranging from optimisation [55, 68], to causal models [25], distributional analyses [9, 22, 33] and the centrality of instances [16, 29, 49, 61]. All of these techniques claim that the counterfactuals they generate are "good" explanations for end-users, although they often differ on how the notion of "goodness" is operationalised, whether by virtue of their "proximity" to the query [68], "plausibility" [25], "sparsity" [29] or "diversity" [55]. However, few of these "goodness" claims have been specifically substantiated in user tests. Indeed, a 2021 review found that just 21% of 117 papers on counterfactual explanation included any form of user-testing, and even fewer (only 7%) tested specific aspects of a proposed method [28]. This absence raises the issue that many current XAI counterfactual explanation techniques lack psychological validity, that is, their explanations may not concretely impact people's understanding of the AI system or its decision, and they may have no practical benefits in real-life applications [2, 41].

Existing studies tend to assess general questions about the impact of counterfactual explanations on users' evaluations. Some

compare people's performance given counterfactual explanations to no-explanation controls [16, 43, 46]. A few studies compare counterfactual explanations to other (*e.g.*, example-based) explanation strategies [65]. Moreover, some studies use *objective* measures (*e.g.*, accuracy of user predictions), whereas others use *subjective* measures (*e.g.*, user judgments of trust, satisfaction, preference). These studies show mixed support for the use of counterfactual explanations in improving user understanding of AI systems. "What-if" counterfactual explanations improved performance in prediction and diagnosis tasks relative to no-explanation controls, but not more so than other explanation options ("why-not", "how-to" and "why" explanations) [43]. Visual counterfactual explanations have been shown to increase classification accuracy relative to no-explanation controls in a small sample of users [16]. However, it has also been found that prompting users to reason counterfactually about a decision can impair objective performance. Lage et al. [37] compared counterfactual tasks, in which users predicted if a system's recommendation would change given a perturbation of some input feature, to simulation tasks, in which users predicted the recommendation based on input features. Counterfactual tasks elicited longer response times, greater judgments of difficulty, and lower accuracy than forward simulation. Another study found that people were less accurate when asked to produce a counterfactual change for an instance than when asked to predict an outcome from the features [46]. These findings are consistent with psychological evidence that counterfactuals aid people to reason about past and future decisions, but in doing so require cognitive effort and resources [6, 47, 48, 59] as we describe later. However, caution needs to be exercised in drawing conclusions from this small collection of studies, given their diversity of tasks, domains and experimental designs, and particularly given their methodological weaknesses, especially given the lack of controls in some of them, the too few test items or very small numbers of participants.

Another set of counterfactual XAI studies have focused on whether counterfactual explanations improve people's self-reported trust or satisfaction in the AI system. They have found generally positive results. Users judge counterfactuals as more appropriate and fair than example-based [3], demographic-based, and influence-based explanations [10]. Contrastive explanations were found to increase self-reported understanding of a system's decisions [46]. Crucially however, some studies show dissociations between objective and subjective measures, that is, users subjectively prefer a certain explanation, but it does not objectively improve their understanding [4, 35]. Users shown contrastive rule-based explanations self-reported better understanding of the system's decision than no-explanation controls, however, neither contrastive rule-based nor contrastive example-based groups differed from a no-explanation group in predictive accuracy for what the system might do, and they tended to follow the system's advice even when incorrect [65]. Thus, studies asking users how well they understand a system's decisions or how satisfying they find an explanation, may not accurately reflect the *actual* impact of an explanation, particularly given people's propensity to overestimate their understanding of complex causal mechanisms [60]. Notably, if an explanation has no objective impact on understanding but is subjectively preferred by users, then concerns about its ethical use arise.

In addition to counterfactuals, the present work also considers the effects of causal explanations on user understanding and judgments in XAI. Causal explanations have a long history in AI, often cast as decision sets and decision trees (*e.g.*, [5, 21, 63]), and typically take the form of "IF-THEN" rule statements consisting of a condition, which if met, leads to a prediction, such as "if the customer's salary is under $10k, then do not grant the loan". Recently, concise decision sets describing local decision-boundaries [39] have been shown to facilitate faster, more accurate user understanding than complete rule lists. Many XAI researchers argue that these rule-based explanations are human-interpretable as post-hoc explanations for opaque models [11, 40, 58], although some have questioned this claim [44]. Recently, some XAI user studies have examined causal rules. For instance, Lage et al. [37] report that decision sets elicit high predictive accuracy, low self-reported difficulty, and quick response times from end-users. The same study found similar divergence between objective and subjective evaluation measures to that identified in counterfactual user studies [35, 65]; as tasks systematically increased the complexity of a system's causal rules, so too subjective judgments of difficulty also increased, as did users' response times; however, little effect of complexity was observed on task accuracy. The present study compares counterfactual and causal explanations for AI decisions, examining their effects on accuracy of understanding as well as on self-reported satisfaction and trust, and deriving hypotheses from extensive psychological research on both, to which we now turn.

2.2 Cognitive Differences Between Counterfactual and Causal Explanations

The present work compares the effects of counterfactual explanations on user understanding to those of causal explanations. Although few XAI studies have compared these two explanation methods, there has been considerable interest in philosophy [20, 42], see also [18], and in psychology [38, 45, 51, 53, 62] in the interdependence between counterfactual and causal reasoning. Reasoning about a counterfactual alternative, in which an outcome would have turned out differently if a preceding event had been different, has been shown to amplify judgments of a causal link between the event and outcome [50]. People's understanding of causality often relies on counterfactual reasoning about different possible outcomes [15]. However, counterfactual and causal reasoning are psychologically distinct. People tend to construct causal explanations that focus on strong causes, *e.g.*, a drunk driver swerving into the protagonist's car caused the crash, *i.e.*, a cause that is necessary and sufficient for the outcome to occur; whereas counterfactual explanations tend to focus on enabling causes, *e.g.*, the crash would have been avoided if the protagonist had taken a different route, *i.e.*, a cause that is necessary but not sufficient for the event to come about [6, 47]. When people create or understand a counterfactual such as "if she had applied for a loan under $5k, it would have been approved" they mentally envisage two possibilities: the conjecture, she applied for a loan under $5k and it was approved, and the presupposed facts, she didn't apply for a loan under $5k and it wasn't approved [6] see also [57]. In contrast, when they understand the causal assertion "the loan wasn't approved because she didn't apply for one under $5k", they envisage only a single possibility initially, corresponding

to the specified facts. Accordingly, the richer mental representation of counterfactuals confers a cognitive benefit, *e.g.*, people make more inferences from counterfactuals than from factual assertions [8]. But the cognitive benefits of counterfactuals come at the cost of requiring more cognitive resources to maintain and update multiple mental representations. Hence, people tend to spontaneously create twice as many causal explanations as counterfactual thoughts when reflecting on an imagined negative event [51]. On the one hand, a counterfactual explanation for an AI system's decision will prompt users to simulate the dual possibilities invoked by comparing the factual input and output of the AI system to a counterfactual case with a different input and output, which may prove effective in improving users' mental models of an AI system. On the other hand, a causal explanation will be less cognitively demanding for users, since it does not require mental simulation of multiple possibilities. The current experiments follow up such work by comparing counterfactual and causal explanations for matched instances, testing whether people find counterfactual explanations more helpful than causal ones, subjectively in self-reports of explanation satisfaction, and objectively in the accuracy of their understanding of the AI system.

2.3 Feature-types and Counterfactual Explanations

The present work also studies the impact of different feature-types on users' understanding and evaluation of an AI system. Many counterfactual explanation algorithms prioritise using specific feature-types in counterfactuals, such as causally-important [25], or actionable (*i.e.*, within the user's control [1, 64]) features. For example, telling a bank customer "if you were to reduce your education level, you would have been granted the loan" is not viewed as useful since lowering one's level of education is not an "actionable" change. Such ideas about the mutability or actionability of features are intuitively appealing and easily implemented in counterfactual methods, but evidence of their psychological impact is less clear-cut. In the only XAI study that has examined actionable features, it was found that actionability predicts user satisfaction and (self-reported) understanding [34]. However, the authors also highlight that ideas of feature mutability and actionability were not easy to define and were highly context-dependent (*e.g.*, increasing one's salary by $1,000 might be highly actionable for a high-earning executive, but entirely non-actionable for someone earning the minimum wage).

Crucially, counterfactual algorithms appear to overlook fundamental distinctions in feature-types that directly impact human understanding. Kahneman & Tversky [24] point out that people do not spontaneously make small changes to continuous variables when creating counterfactuals (*e.g.*, they do not say "if only the driver had driven through the junction two seconds earlier, the accident could have been avoided"). People may not tend to change continuous features in counterfactuals, perhaps because they find them harder to identify or understand. For example, a counterfactual explanation that tells a user, "if your credit score had been high, your loan application would have been approved" may be easier to understand than one saying "if your credit score had been 4.6, your loan application would have been approved". The first provides a binary distinction, between the categorical features, high and low, whereas the second provides a specific point on a continuous scale. People find it easier to reason about binary alternates rather than a contrast class consisting of multiple values [13, 56]. However, counterfactual algorithms do not take into account this distinction. Some methods suggest meaningless non-integer values for categorical features (*e.g.*, Race= .5 [68]), whereas others use one-hot encoding to transform categorical features into continuous variables [25, 55] or project categorical variables onto an ordinal feature-space [9, 66]. These methods focus on transforming categorical features into continuous formats, implicitly assuming that continuous and categorical features are interchangeable, when in fact, people may understand explanations based on categorical and continuous features very differently. In the present studies, we compare the effects of explanations focusing on continuous and categorical features on people's understanding of an AI system's decisions. We hypothesise that categorical features will be more readily understood than continuous features.

2.4 Outline of Paper

In the remainder of this paper, we report two experiments that compare the impact of counterfactual and causal explanations, and the impact of explanations based on continuous and categorical features, on users' understanding and subjective evaluation of a simulated AI system designed to predict BAC thresholds. Participants were shown predictions by the system for different instances, accompanied by explanations. The explanations were phrased as either counterfactual or causal assertions, and they were about either continuous or categorical features. Participants gained experience of the system's predictions and learned about the BAC domain with the help of the explanations. Then, participants' understanding of the system was objectively measured using the accuracy of their predictions, without feedback or explanations. Finally, users' subjective evaluations, in the form of explanation satisfaction and trust, were recorded.

3 EXPERIMENTAL TASK: AN APP THAT PREDICTS LEGAL DRIVING LIMITS

Participants were presented with alcohol driving-limit decisions from a simulated automated system application; the application was presented as designed to predict whether someone was over the legal BAC limit to drive. The decisions were based on a commonly-used approximate method, the Widmark equation [70], that uses five features to assess BAC, with the limit threshold set at 0.08% alcohol per 100ml of blood. This formula was used to generate a dataset of instances for normally-distributed values of the feature-set (N=2000), from which the study's materials were drawn.

In the experimental task, participants were instructed that they would test a new application, *SafeLimit*, designed to inform people whether or not they are over the legal limit to drive, based on five features: *units* of alcohol consumed by the person, *weight* (in kg), *duration* of drinking period (in minutes), *gender* (male/female) and *stomach-fullness* (full/empty). The experiment consisted of two phases: a *training phase*, in which they made predictions and were given explanations and feedback on their responses, and a *testing phase*, in which they made predictions for instances without explanations or feedback. These phases tested whether experience of

Figure 1: Sample item from the training phase in the Counterfactual condition of Experiment 1. In (a) is an example of a single trial with the three response options, in (b) is the feedback shown to a participant who chose the correct answer (in this case "over the limit"), including the explanation, in (c) is the feedback shown to a participant who chose the incorrect answer.

the app's predictions with/without explanations impacted people's knowledge of the application and domain.

In the training phase, participants were shown tabular data for different individuals on each screen and asked to make a judgment about whether each individual was under or over the limit on each screen. Participants selected one of three options: "Over the limit", "Under the limit", or "Don't know" by clicking the corresponding on-screen button (see Figure 1a). The order of these options was randomised, to ensure that participants did not click on the same button placement each time. After the participant had submitted a response, they were shown feedback on the next screen, that is, their chosen answer was highlighted with a green tick-mark if it was correct (see Figure 1b), or a red X-mark if it was incorrect (see Figure 1c) and the correct answer was highlighted in green. In addition to the feedback, participants were also shown an explanation on this screen, placed above the answer options (see Figure 1b and Figure 1c). The explanation they were shown depended on the experimental condition to which they were assigned, counterfactual, causal or no explanation. Figure 1 shows sample materials used in the counterfactual explanation condition.

The testing phase started after completing the training phase. Here, participants were shown instances describing different individuals (see Figure 2) and asked to judge if the individual was over the legal limit to drive. After submitting each response, no feedback or explanation was given, and they moved on to the next trial. For each instance, participants were asked to consider a specific feature in making their prediction, *e.g.*, "Given this person's WEIGHT, please make a judgment about their blood alcohol level."

The objective measure of performance in the study was accuracy (*i.e.*, correct predictions made by participants, that is, whether the participant's prediction of the AI's output aligned accurately with the output that the AI system would actually make). The subjective measures were explanation satisfaction and trust in the system, assessed using DARPA's Explanation Satisfaction and Trust scales [19], respectively (see appendix A.1 and A.2). Participants were provided with general information about the experiment at the outset and they were invited to provide their consent to participate. Participants who consented to take part in the experiment read detailed instructions about the task and completed practice trials for each phase of the study before commencing. They progressed through the presented instances in a different randomised order for each participant, within the training and testing phases. After completing both phases, they completed the Explanation Satisfaction and Trust scales, and were debriefed and paid for their time. To ensure participants included in the analysis had engaged with the task, all participants completed four attention checks at random intervals throughout the experiment, and at the end of the session, they were asked to recall the 5 features used by the application by selecting them from a list of 10 options. The task instructions and data for the experiments are available at https://osf.io/dmvc2/. Ethics

approval for both studies was granted in advance by University College Dublin with the reference code LS-E-20-11-Warren-Keane.

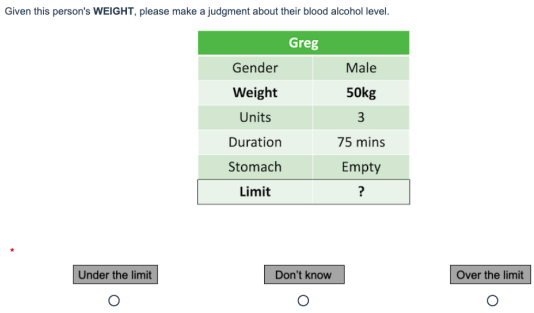

Figure 2: Sample item from the testing phase of Experiment 1

4 EXPERIMENT 1

The aim of the experiment was to compare the effect of counterfactual and causal explanations for the *SafeLimit* application's decisions to no-explanation controls (which re-described the decision without explanation). Participants were assigned to one of three groups (counterfactual, causal, or no-explanation control) and completed the (i) training phase in which they made predictions and were given feedback with explanations (explanation groups) or re-descriptions (control group) and (ii) testing phase where they made predictions with no feedback or explanations (for all groups). Hence, any observed differences in accuracy in the testing phase will reflect participants' experience of the explanations or no-explanation in the training phase, given that this variation was the sole difference between conditions. Participants were presented with 40 items in each phase, which systematically varied in terms of the five features used with balanced occurrence (*i.e.*, eight instances for each feature). Explanation satisfaction and trust in the system were measured following the training and testing phases. Our primary predictions were: (i) explanations will improve accuracy, that is, performance in the testing phase will be more accurate than performance in the training phase, (ii) counterfactual explanations will improve accuracy more than causal explanations, if they are more informative, (iii) predictions about categorical features will be more accurate than predictions about continuous features, if people find the former easier to understand than the latter, and (iv) counterfactual explanations will be judged as more satisfying and trustworthy than causal explanations, replicating previous studies showing that they are often subjectively preferred over other explanations.

4.1 Method

4.1.1 Participants and Design. The participants (N=127) were recruited using Prolific (https://www.prolific.co/) and they were assigned in fixed order to the three between-participant conditions: counterfactual explanation (n=41), causal explanation (n=43) and no-explanation control (n=43). The participants comprised 80 women, 46 men, and one non-binary person and they were aged 18-74 years (M=33.54, SD=13.15). Participants were pre-screened to be native English speakers from Ireland, the United Kingdom, the United States, Australia, Canada and New Zealand, who had not participated in previous related studies. A further 11 participants were excluded prior to data analysis, one for giving identical responses for each trial, and 10 who failed more than one attention or memory check. The experimental design was a 3 (Explanation: counterfactual, causal, control) x 2 (Task: training vs testing phase) x 5 (Feature: units, duration, gender, weight, stomach-fullness) design, with repeated measures on the latter two variables. Before testing, the power analysis with G*Power [14] indicated that 126 participants were required to achieve 90% power for a medium-sized effect with alpha <.05 for two-tailed tests.

4.1.2 Materials and Procedure. Eighty instances were randomly selected from the 2000-item dataset generated for the BAC domain. Specifically, the procedure randomly selected an instance (query case) and incrementally perturbed one of the five feature values until its BAC value crossed the decision-boundary to create a counterfactual case. If the perturbation was successful, the instance was selected as a material and its counterfactual was used as the basis for the explanation shown to the counterfactual group. For example, if an instance with *units*=4 crossed the decision-boundary when it was reduced by one unit (to be under rather than over the limit), the counterfactual explanation read "If John had drunk 3 units instead of 4 units, he would have been under the limit"; the matched causal explanation read "John is over the limit because he drank 4 units"; and the control group was given a re-description of the outcome "John is over the limit". This selection procedure was performed 16 times for each feature, a total of 80 times, with balanced instances on either side of the decision-boundary (*i.e.*, equal numbers under and over the limit). Each instance was randomly assigned to one of two sets of materials, each comprising 40 items, again ensuring a balanced number of instances. To avoid any material-specific confounds, materials presented in the training and testing phases were counterbalanced, so that half of the participants in each group saw Set A in the training phase, and Set B in the testing phase, and this order was reversed for the other half of the participants. After data collection, t-tests verified that there was no effect of material-set order. Participants were paid £2.61 for their time. The experiment took approximately 28 minutes to complete.

4.2 Results

The results show that providing explanations improved the accuracy of people's predictions in the testing phase; participants judged counterfactual explanations to be more satisfying and trustworthy than causal explanations, but counterfactual explanations had only a slightly greater impact than causal explanations on participants' accuracy in predicting the app's decisions. Categorical features led to higher prediction accuracy than continuous features. Participants' accuracy on categorical features was markedly higher in the testing phase than in the training phase, whereas their accuracy on continuous features remained at similar levels in both phases (an effect that occurred independently of the explanation type).

4.2.1 User Accuracy. To test the hypotheses, a 3 (Explanation: counterfactual, causal, control) x 2 (Task: training vs testing) x 5 (Feature: units, duration, gender, weight, stomach-fullness) mixed ANOVA with repeated measures on the second two factors was conducted

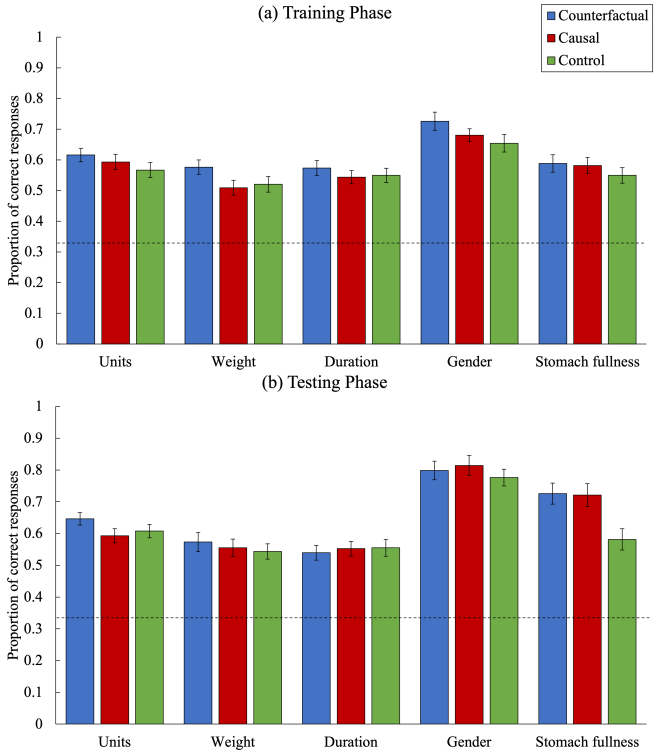

Figure 3: Task accuracy (proportion of correct responses) for the three conditions in Experiment 1 for each of the 5 features in the (a) training and (b) testing phases (error bars show standard error of the mean; the dashed line represents chance accuracy).

on the proportion of correct answers given by each participant (and the mean proportions are provided in Figure 3). A significant main effect was found for Explanation, $F(2,124)=5.63$, $p=.005$, $\eta_p^2=.083$, and post hoc Tukey HSD tests showed that the Counterfactual group ($M=.636$, $SD=.08$) was more accurate than the Control group ($M=.590$, $SD=.08$), $p=.003$, $d=.22$. The Causal group ($M=.614$, $SD=.09$) was not significantly more accurate than the Control group, $p=.186$; or the Counterfactual group, $p=.245$. These results indicate that providing explanations is better than not providing them, for improving accuracy. They also show, as predicted, that counterfactual explanations have a greater effect on accuracy than causal explanations or no-explanation controls. Note that these effects were observed for both phases of the study overall (Explanation did not interact with Task).

There were also main effects for Task, $F(1,124)=32.349$, $p<.001$, $\eta_p^2=.207$, and for Feature, $F(3.945, 489.156)=47.599$, $p<.001$, $\eta_p^2=.277$, and Task interacted with Feature, $F(4, 496)=7.23$, $p<.001$, $\eta_p^2=.055$.[1] No other effects were significant.[2] Each of the significant effects were further examined in post hoc analyses. The decomposition of the interaction revealed that accuracy improved from the training to the testing phase for the categorical features (*gender*, *stomach-fullness*), but not for the continuous features (*units*, *weight* and *duration*). Post hoc pairwise comparisons with a Bonferroni-corrected alpha of .002 for 25 comparisons showed that participants made more correct responses in the testing phase than in the training phase when considering *gender*, $t(126)=5.626$, $p<.001$, $d=.50$, and *stomach-fullness*, $t(126)=4.430$, $p<.001$, $d=.39$, but not *units*, $t(126)=1.350$, $p=.179$, *weight*, $t(126)=-1.209$, $p=.229$, or *duration*, $t(126)=.32$, $p=.75$. The analysis also showed that within each phase of the study, the categorical features produced higher accuracy than the continuous features, confirming the prediction that people find the former easier to understand than the latter. In the training phase, accuracy for *gender* was significantly higher than accuracy for *units*, $t(126)=4.935$, $p<.001$, $d=.44$, *weight*, $t(126)=6.824$, $p<.001$, $d=.61$, *duration*, $t(126)=6.332$, $p<.001$, $d=.58$, and *stomach-fullness*, $t(126)=5.202$, $p<.001$, $d=.46$, the other features did not differ significantly from each other ($p>.05$ for all comparisons). In the testing phase, accuracy was higher for *gender* than for *units*, $t(126)=8.844$, $p<.001$, $d=.78$, *weight*, $t(126)=10.824$, $p<.001$, $d=.96$, *duration*, $t(126)=10.81$, $p<.001$, $d=.96$ and *stomach-fullness*, $t(126)=4.986$, $p<.001$, $d=.44$. Accuracy for *stomach-fullness* was higher than for *weight*, $t(126)=4.943$, $p<.001$, $d=.44$, *duration*, $t(126)=4.959$, $p<.001$, $d=.44$, and *units*, $t(126)=2.853$, $p=.005$, although the latter was not significant on the corrected alpha.[3]

4.2.2 User Accuracy: exploratory analysis. To probe the nature of the significant effects of explanations further we carried out an exploratory analysis, that is, an analysis that had not formed part of our initial hypotheses. It indicated a reliable trend of increasing accuracy in group scores in the following order: Counterfactual > Causal > Control, Page's $L(40)=1005.0$, $p<.001$ (the median scores are provided in Figure 4a).

We carried out a second exploratory analysis, to probe further the effects of features. It indicated it is the diversity in the range of feature-values that likely leads to their effects, rather than some abstract ontological status of the feature. When we rank-ordered each of the features in terms of the number of unique values present in the materials, we found that it predicted the observed trend in accuracy in the testing phase. That is, the rank ordering from highest-to-lowest diversity – *duration* (60 unique values) > *weight* (36 unique values) > *units* (4 unique values) > *stomach-fullness* (2 unique values) = *gender* (2 unique values) – inversely predicts the trend in accuracy: *duration* ($M=.549$) < *weight* ($M=.557$) < *units* ($M=.615$) < *stomach-fullness* ($M=.675$) < *gender* ($M=.796$); Page's $L(127)=6256.5$, $p<.001$ (the median scores are provided in Figure 4b).

4.2.3 Subjective Evaluation. **Explanation Satisfaction Measure.** A one-way ANOVA was carried out on the summed judgments for the Explanation Satisfaction scale to examine group differences in satisfaction levels for the explanations provided (the summed judgment scores are provided in Figure 5a). Significant differences between the groups were identified $F(2, 126)=6.104$, $p=.003$, $\eta_p^2=.09$.

[1] A Huynh-Feldt correction was applied to the main effect of Feature and its interactions.
[2] No other two-way interactions were reliable, Explanation did not interact with Task, $F(2, 124)=.759$, $p=.47$, nor with Feature, $F(7.89, 489.156)=1.14$, $p=.335$, and the three variables did not interact, $F(8, 496)=1.215$, $p=.288$.

[3] Accuracy for *units* was significantly higher than *weight*, $t(126)=3.152$, $p=.002$, $d=.28$ and *duration*, $t(126)=3.539$, $p=.001$, $d=.31$. Accuracy for *weight* and *duration* did not differ, $t(126)=.385$, $p=.701$.

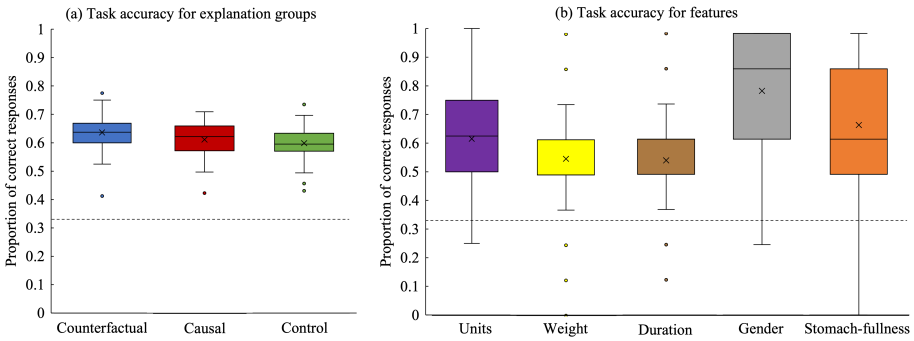

Figure 4: Task accuracy (proportion of correct responses) for (a) the three explanation conditions in Experiment 1, (b) for each feature in Experiment 1 (lines within boxes denote medians, x-marks denote means, whiskers represent 1.5 times the interquartile range, dots represent outliers, and the dashed line represents chance accuracy).

Post hoc Tukey HSD tests showed that participants in the counterfactual group (M=27.83, SD=6.12) self-reported significantly higher satisfaction with the explanations than those in the causal group (M=22.79, SD=6.63), p=.002, d=0.76. The control group (M=25.86, SD=7.19) did not differ significantly from either the counterfactual (p=.369) or the causal (p=.087) groups.

Trust Measure. A one-way ANOVA was carried out on the summed judgments for the Trust Scale to examine group differences in trust levels for the explanations provided (see Figure 5b). Significant differences between the groups were identified $F(2, 126)$=8.184, p<.001, η_p^2=.117. Post hoc Tukey HSD tests showed that participants in the counterfactual group (M=26.15, SD=6.14) self-reported significantly higher trust in the AI system than those in the causal group (M=20.21, SD=6.27), p<.001, d=.88. The control group (M=23.12, SD=7.63) did not differ significantly from either the counterfactual (p=.101) or causal groups (p=.115).

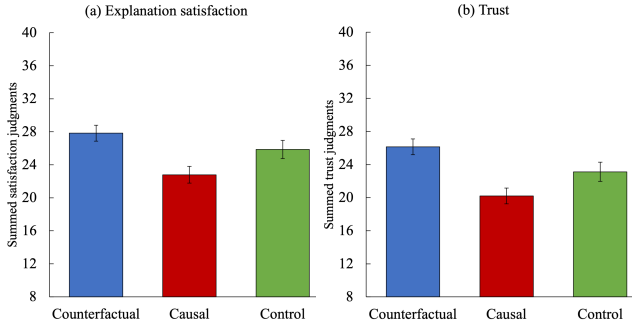

Figure 5: Summed judgment scores for (a) Explanation satisfaction and (b) Trust in Experiment 1 (minimum score possible is 8, maximum is 40, error bars represent standard error of the mean).

4.2.4 Subjective Evaluation: exploratory analysis. We also carried out an exploratory analysis on the subjective evaluations. For the explanation satisfaction measure, a reliable trend was identified when rank-ordering judgments for each item in the order: Counterfactual > Control > Causal, Page's L(8)=111.0, p<.001, suggesting that counterfactual explanations were somewhat more satisfying than non-explanations, and non-explanations were somewhat more satisfying than causal explanations. People were less satisfied with causal explanations compared to counterfactual explanations.

For the trust measure, a reliable trend was identified when rank-ordering judgments for each item in the order: Counterfactual > Control > Causal, Page's L(8)=112.0, p<.001. Like the satisfaction judgments, these results suggest that counterfactual explanations were judged somewhat more trustworthy than non-explanations, and non-explanations were somewhat more trustworthy than causal explanations. People placed less trust in causal explanations compared to counterfactual explanations.

4.3 Discussion

Experiment 1 corroborated our primary predictions, showing that (i) explanations improved accuracy; performance in the testing phase was more accurate than performance in the training phase, (ii) counterfactual explanations improved accuracy more than causal explanations; users' prediction accuracy improved when given counterfactual explanations, relative to causal explanations, which in turn were more effective than control descriptions, (iii) participants' predictions about categorical features were more accurate than their predictions about continuous features, and (iv) users who were shown counterfactual explanations gave higher subjective judgments for explanation satisfaction and trust than those who were shown causal explanations, whereas the judgments of no-explanation controls lay in-between those of the counterfactual and causal groups.

4.3.1 Counterfactual and causal explanations. These results show that counterfactual explanations had a greater impact on people's accuracy in understanding an AI system, as well as eliciting higher satisfaction and trust judgments, significantly higher than causal explanations, and somewhat higher than non-explanations. The findings for subjective judgments are consistent with previous findings showing that counterfactual explanations tend to be perceived positively by users [3, 10, 46], as well as bolstering claims that they improve user understanding [7, 54]. However, it is worth noting that users' accuracy scores do not completely align with their subjective evaluation scores. The Counterfactual group made more accurate

predictions than the Control no-explanation group, whereas the Causal group did not; and the trend in accuracy showed that predictions by the Counterfactual group were more accurate than those by the Causal group, which were more accurate than those by the Control no-explanation group. Although participants in the Causal explanation group were more accurate in predicting the system's decisions than the Control group, the Control group gave more favourable satisfaction and trust judgments than the Causal group. This highlights the need for caution in analysing users' self-reported evaluations of explanations, particularly given that people tend to overestimate the depth of their understanding of complex phenomena [60].

It is important to note that the better performance in accuracy, satisfaction, and trust elicited by counterfactual explanations relative to causal explanations may be due to the provision of more information about the relationship between a given feature and the outcome (*i.e.*, how much that feature must change in order to alter the system's prediction), *e.g.*, "if John had drunk 3 units instead of 4 units, he would have been under the limit", compared to the causal explanation, "John is over the limit because he drank 4 units". Although this additional information may be viewed as an inherent benefit of counterfactual explanation over more factual explanations [54], such as causal explanations, feature-importance scores [58] and decision-sets [39], it raises the question of whether these alternatives could be equally effective as their counterfactual counterparts if they contained supplementary information about the decision-boundary. As discussed in section 2.2, if counterfactual explanations prompt users to represent the factual case as well as the counterfactual case, it may be the simulation of both possibilities that aids them in developing a more accurate model of the underlying AI system, rather than merely the additional contrastive information alone. To test this, in the next experiment we matched the amount of information provided by the counterfactual and causal explanations. For example, a counterfactual explanation "if Sarah had weighed 15kg heavier, she would have been under the limit", and a matched causal explanation, "because Sarah's weight was 15kg too light, she was over the limit", provide the same information about the relationship between the feature (*weight*) and the outcome (being over the limit). The causal explanation now also provides contrastive information implicitly, as does the counterfactual explanation. If counterfactual explanations induce users to consider multiple possibilities, we expect that participants shown counterfactual explanations will again exhibit higher accuracy in their predictions of the system's decisions, compared to causal explanations and control non-explanations.

4.3.2 Categorical and continuous feature-types in explanation. We found that users were more accurate in predicting the system's decisions based on categorical features rather than continuous ones. While task accuracy increased from the training phase to the testing phase, this improvement was seen only in accuracy for the categorical features (*gender* and *stomach-fullness*), whereas accuracy for continuous features (*units*, *duration*, and *weight*) remained the same in both phases. This psychological difference between categorical and continuous feature-types has been overlooked by current counterfactual algorithms in XAI, which tend to treat continuous and categorical features as interchangeable (see [69] for our model that tries to take these findings into account). However, the exact source of these benefits for categorical features over continuous features is still unclear. One conjecture is that categorical features have some intrinsic ontological property that makes them easier for people to process. Another conjecture is that categorical features are easier to process because of the lack of diversity in the feature-value-ranges presented to the people; that is, a continuous feature such as weight was given 36 different feature-values over presented instances in the experiment, whereas a categorical feature such as gender was given only 2 different feature-values in the experiment. If the latter conjecture true, then converting continuous features into categorical ones with less feature-value diversity (*e.g.*, discretising the weight feature to appear as heavy/light, that is, to have only 2 different feature-values) should change the lower accuracy observed for continuous features here. Furthermore, although users appear to understand categorical features more readily than continuous ones, it was not possible to test any potential differences in subjective judgments of satisfaction and trust that may result from continuous versus categorical features, due to the within-participants design of the feature-type factor in Experiment 1. In the next experiment, we examine feature-type as a between-participants factor to enable us to do so.

5 EXPERIMENT 2

In Experiment 2, we compared the impact, on users' objective accuracy and their subjective self-reported satisfaction and trust, of (i) mixed features (*i.e.*, categorical and continuous, as presented in Experiment 1) versus categorical features (*i.e.*, all continuous features converted to categorical ones, alongside the existing categorical ones) and (ii) counterfactual explanations, causal explanations, and no-explanation. As in Experiment 1, participants completed (i) a training phase, (ii) a testing phase during which their accuracy was measured, and (iii) self-reported judgments of explanation satisfaction and trust. Participants were presented with 16 items in each phase, which were systematically varied across four of the features with balanced occurrence (*i.e.*, 4 instances for each feature). Our predictions were that (i) users will be more accurate when shown items with only categorical features compared to items with mixed continuous and categorical features, (ii) users presented with counterfactual explanations will be more accurate than those presented with causal explanations or no-explanation controls, and (iii) counterfactual explanations will elicit higher subjective satisfaction and trust than causal explanations and controls, *i.e.*, effects from Experiment 1 will be replicated.

5.1 Method

5.1.1 Participants and Design. The participants (N=211) were subject to the same pre-screening criteria as Experiment 1, and assigned in a fixed order to six groups: mixed-counterfactual explanation (n=34), mixed-causal explanation (n=31), mixed-control (n=37), categorical-counterfactual explanation (n=41), categorical-causal explanation (n=34), and categorical-control (n=34). Participants comprised 100 women, 99 men, three non-binary people, and nine participants who did not disclose their gender or age. Those participants who reported demographic information were aged 18-75 years (M=33.1, SD=11.0). Prior to analysis, 34 participant responses

were excluded because they failed more than one attention or memory check. The experimental design was a 3 (Explanation: counterfactual, causal, control) x 2 (Feature-format: mixed vs categorical) x 2 (Task: training vs testing phase) design, with repeated measures on the third variable. A power analysis conducted with the easypower package for R [52] indicated that 211 participants were required for 90% power, given a medium effect size with alpha <.05 for two-tailed tests.

5.1.2 Materials & Procedure. Participants were presented with 32 instances (16 of which were unique) drawn from a BAC dataset. Participants in the mixed feature groups were shown instances with features in the same format as the participants in Experiment 1, that is, three continuous features (*units*, *weight*, and *duration*) and two binary categorical features (*gender* and *stomach-fullness*). Participants in the categorical feature groups were shown the same instances, but the features of *units*, *weight*, and *duration* were re-coded to be categorical, so that all five features had binary categorical values. The continuous features were coded as categorical by generating a normally-distributed dataset of instances (N=2000) and computing the upper and lower quartile values for each feature (see appendix B.1). *Units*, *weight*, and *duration* were re-coded as 'high'/'low', 'heavy'/'light', and 'long'/'short' respectively. For example, a case with the values *gender* = male, *units* > 5, *weight* > 83kg, *duration* > 106 mins, *stomach-fullness* = empty was re-coded as *gender* = male, *units* = high, *weight* = heavy, *duration* = long, *stomach-fullness* = empty. Instances containing any features with a value in the interquartile range were considered ineligible for selection. Sixteen of the 32 possible combinations of the five categorical features were selected as instances to present to participants, based on their proximity to the decision-boundary. Each instance was presented twice: once in the training phase, and once in the testing phase under a different individual's name. Eight of the instances were predicted to be over the limit, and eight were predicted to be under the limit. In order to ensure each feature was referred to an equal number of times, participants were asked to consider only four of the five features (*gender*, *units*, *weight* and *stomach-fullness*). Each of these features was referred to four times in the explanations in the training phase, and four times in the prompts in the testing phase *i.e.*, for 16 instances in each phase. The remaining feature of *duration* was presented with the other features in each trial, however it was not referred to in any of the explanations or prompts.

The explanations presented in the training phase referred to a continuous or categorical change in the features, depending on the experimental condition. The wording of the counterfactual and causal explanations was adjusted from Experiment 1 to provide matched information about how a feature would have to change to alter the system's decision (see appendix B.2 for examples). In the instructions, participants were told how the continuous features had been categorised into categorical ones and reminded of this information before beginning the testing phase. Upon completion, participants were compensated £1.40. The experiment took an average time of 20 minutes to complete.

5.2 Results

The results of Experiment 2 show that providing explanations and presenting features as categorical led to higher user accuracy in predicting the app's decisions. Participants presented with instances involving only categorical features were more accurate in predicting the app's decision based on these features than participants shown mixed continuous and categorical features. Participants given counterfactual explanations were more accurate compared to the causal explanation and control groups, regardless of feature-format. Participants given control non-explanations gave higher judgments of satisfaction and trust than those in the explanation conditions and participants shown mixed features placed more trust in the app than those shown only categorical features.

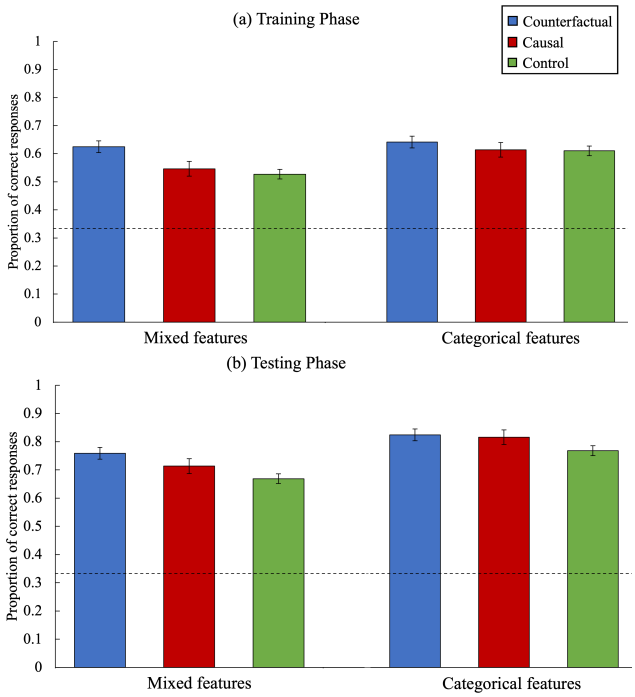

Figure 6: Task accuracy for the six conditions in Experiment 2 in the (a) training and (b) testing phase (error bars represent standard error of the mean, dashed line represents chance accuracy).

5.2.1 User Accuracy. A 3 (Explanation: counterfactual, causal, control) x 2 (Feature-format: mixed vs categorical) x 2 (Task: training vs testing) mixed ANOVA with repeated measures on the third factor was conducted on the proportion of correct responses given by each participant (see Figure 6). There was a main effect of Explanation, $F(1, 205)=6.839$, $p=.001$, $\eta_p^2=.063$, and post hoc Tukey HSD tests indicated that the Counterfactual explanation group ($M=.72$, $SD=.16$) was significantly more accurate than the Control group ($M=.64$, $SD=.16$), $p<.001$, $d=.64$, while the Causal group ($M=.68$, $SD=.19$) did not differ significantly from the Counterfactual group, $p=.094$, or Control group, $p=.211$. There was a main effect of Feature-format, $F(1, 205)=21.558$, $p<.001$, $\eta_p^2=.095$, as the categorical feature group ($M=.71$, $SD=.17$) was significantly more accurate than the mixed feature group ($M=.64$, $SD=.16$), *i.e.*, participants presented with only categorical features were more accurate than those shown items involving a mix of continuous and categorical features, regardless of the explanation they received. There was also a main effect of

Task, $F(2, 205)=193.762$, $p<.001$, $\eta_p^2=.486$, as participants were more accurate in the testing phase ($M=.76$, $SD=.16$) than the training phase ($M=.60$, $SD=.13$). No other effects were significant.[4]

5.2.2 User Accuracy: exploratory analysis. Once again, we carried out an exploratory analysis to examine the effect of explanations further. It identified a reliable trend in group accuracy scores in the following order: Counterfactual > Causal > Control, Page's $L(32)=415.0$, $p<.001$ (the median scores are provided in Figure 7). These results show that providing explanations increased prediction accuracy compared to not providing them, with a greater effect of counterfactual explanations on accuracy relative to causal explanations and no-explanation controls.

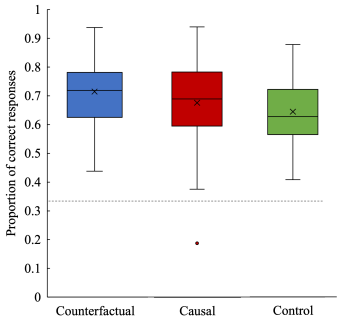

Figure 7: Task accuracy for the three explanation conditions in Experiment 2 (lines in boxes denote medians, x-marks denote means, whiskers show 1.5 times interquartile range, dots show outliers, dashed line represents chance accuracy).

5.2.3 Subjective Evaluation. **Explanation Satisfaction Measure.** A 3 (Explanation: counterfactual, causal, control) x 2 (Feature-format: mixed vs categorical) ANOVA was performed on the summed judgments for the Explanation Satisfaction Scale (see Figure 8a). A main effect of Explanation was found, $F(2, 205)=6.17$, $p=.003$, $\eta_p^2=.057$, no effect of Feature-format, $F(1, 205)=.153$, $p=.696$, and no interaction between factors, $F(2, 205)=1.174$, $p=.311$. Post hoc Tukey HSD tests indicated that the Control group ($M=29.86$, $SD=6.54$) self-reported higher satisfaction with explanations than the Counterfactual group ($M=26.73$, $SD=6.85$), $p=.019$, $d=.45$, and the Causal group ($M=25.90$, $SD=7.59$), $p=.004$, $d=.56$. The Counterfactual group did not differ significantly from the Causal group, $p=.801$.

Trust Measure. A 3 (Explanation: counterfactual, causal, control) x 2 (Feature-format: mixed vs categorical) ANOVA was conducted on summed judgments for the Trust Scale (see Figure 8b). A main effect of Explanation was identified, $F(2, 205)=4.554$, $p=.012$, $\eta_p^2=.043$, and a main effect of Feature-format, $F(1, 205)=5.927$, $p=.016$, $\eta_p^2=.028$. The two factors did not interact, $F(2, 205)=.660$, $p=.518$. Post hoc Tukey HSD tests showed that participants in the Control group ($M=26.56$, $SD=6.70$) self-reported significantly higher trust in the AI system than those in the Counterfactual group ($M=23.83$, $SD=7.04$), $p=.039$, $d=.41$ and Causal group ($M=23.18$, $SD=7.05$), $p=.011$, $d=.50$. The counterfactual and causal groups did not differ in judgments of trust, $p=.844$. The main effect of Feature-format arises because the mixed group ($M=25.85$, $SD=6.98$) gave higher trust judgments than the categorical group ($M=23.36$, $SD=6.88$), indicating that people placed more trust in the AI system when shown mixed features compared to only categorical features.

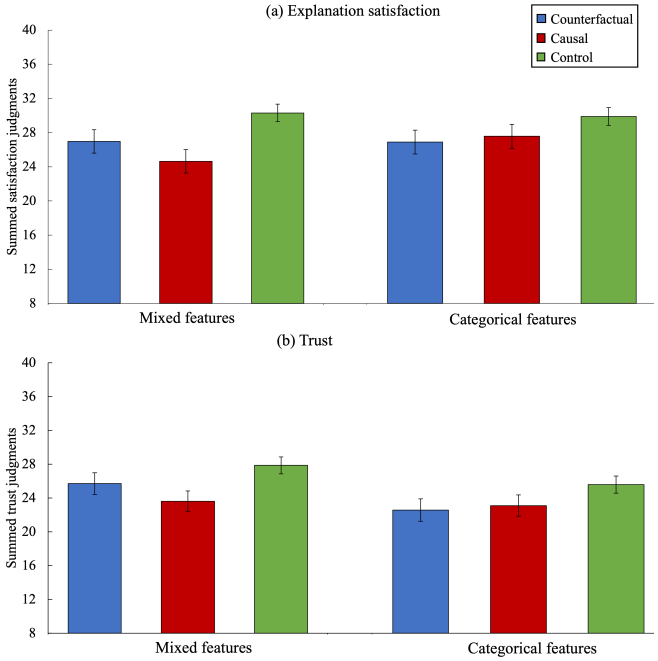

Figure 8: Summed judgments for (a) Explanation satisfaction and (b) Trust in Experiment 2 (minimum score possible is 8, maximum is 40, error bars show standard error of the mean).

5.2.4 Subjective Evaluation: exploratory analysis. In a further exploratory analysis on the explanation satisfaction measure, a reliable trend was identified when rank-ordering judgments for each item in the order: Control > Counterfactual > Causal, Page's $L(8)=110.0$, $p<.001$, indicating that no-explanations were more satisfying than counterfactual explanations, and counterfactual explanations were more satisfying than causal explanations. A reliable trend was also identified for the trust measure when rank-ordering judgments for each item in the order: Control > Counterfactual > Causal, Page's $L(8)=111.0$, $p<.001$, indicating that no-explanations were judged more trustworthy than counterfactual explanations, and counterfactual explanations were judged more trustworthy than causal explanations.

5.3 Discussion

Experiment 2 supported two of our predictions, showing that users of an AI application were more accurate in predicting its decisions when shown, (i) categorical features only, compared to items that mix continuous and categorical features, and (ii) counterfactual explanations compared to causal explanations or no-explanation

[4]The three factors did not interact, $F(2, 205)=.165$, $p=.848$, nor did Explanation interact with Feature-format $F(2, 205)=1.061$, $p=.348$, or task, $F(2, 205)=.754$, $p=.472$. There was no interaction between Feature-format and Task, $F(1, 205)=1.982$, $p=.161$.

at all. The main effect of task suggests that all participants improved after the training phase, regardless of explanation strategy. Nonetheless, the main effect of explanation shows that providing counterfactual explanations had added benefits over causal explanations and no-explanation, and the main effect of feature-format indicates that presenting all the item's features in a categorical format had added impact relative to the other conditions. However, our third prediction regarding users' subjective judgments was not supported: unexpectedly and contrary to the results of the previous experiment, users judged no-explanation control descriptions more favourably than counterfactual and causal explanations.

5.3.1 Counterfactual and causal explanations. Experiment 2 replicates the findings of Experiment 1, in that participants given counterfactual explanations were more accurate than those given no-explanations, and somewhat more accurate than those given causal explanations. This finding was irrespective of whether the items being explained involved only categorical features or mixed continuous and categorical features. As in Experiment 1, participants' accuracy improved from the training phase to the testing phase, regardless of the explanations. Strikingly, users shown counterfactual explanations were reliably more accurate than those shown causal explanations, despite these explanations containing matched information about the decision-boundary. This result suggests that counterfactual explanations have additional explanatory benefits, that are not solely due to providing additional information to the user.

However, the effects for the subjective measures found in Experiment 1 were not replicated. Instead, the no-explanation controls showed higher levels of satisfaction and trust than either of the explanation groups. The trend for satisfaction and trust of Control > Counterfactual > Causal indicates that counterfactual explanations were evaluated as higher in satisfaction and trust than causal ones, as found in Experiment 1; however no-explanation control descriptions were evaluated as higher in satisfaction and trust than either. Of course in both experiments, the subjective judgments were made in the between-participants design of the present studies, that is, participants did not compare counterfactual, causal, and no-explanation descriptions, they assessed only the information given in their allocated condition. Participants in the control groups may have made their subjective judgments relative to their assessment of the *SafeLimit* app as a whole, rather than evaluating the system's explanations. Had our experiments employed a within-participants design, or enforced a choice between explanation types, we would not anticipate a control description to be preferred to counterfactual or causal explanation. The counterfactual and causal explanations in Experiment 2 were modified to ensure that the causal explanations contained comparable contrastive information to the counterfactual ones, and so a possible explanation for the result is that the no-explanation control descriptions, which merely restated explicitly the outcome, appeared simpler than the counterfactual or causal explanations. Irrespective of the exact reasons for this finding, it is somewhat worrying that those participants with the poorest understanding of the system (*i.e.*, the control group has the lowest accuracy scores) gave the most positive judgments of it (the control group had the highest satisfaction and trust scores).

5.3.2 Categorical and continuous features in explanations. Users who were presented with items using only categorical features were more accurate in predicting the system's decisions than those shown the same information in a raw, continuous format. These results confirm our hypothesis that participants understand categorical features more readily than continuous ones. Moreover, they indicate that continuous features, transformed into categorical features (in this case, *units*, *weight* and *duration*) are as easily understood as features that are categorical in their raw form (*e.g., gender* and *stomach-fullness*). In terms of participants' satisfaction, no difference between feature-formats emerged, suggesting that users perceived them to be similarly effective in explanations. However, a difference in user trust did emerge. Users shown only categorical features reported lower levels of trust in the AI system than those shown continuous feature-values. This result may align with a recent study that found that observers of an AI agent ascribed it more intelligent, higher-order thinking when it gave explanations containing numerical data as opposed to natural language, even when the figures were not meaningful [12]. Participants presented with the raw, continuous features may have perceived the system as more precise, and hence more trustworthy, while in contrast, the transformation of continuous features may have been regarded as less reliable and transparent. This finding also indicates a further discrepancy between objective and subjective measures; the participants shown only categorical features were more accurate in predicting the app's decisions, however, they appeared to trust the system less than their counterparts who were shown mixed continuous and categorical features. These results demonstrate that this divergence in measures occurs not only between different sorts of explanations (counterfactual and causal) but also for different feature-formats, again highlighting the importance of measuring both objective and subjective criteria in any user evaluations.

6 GENERAL DISCUSSION

The present studies report key findings for XAI on how explanations impact people's knowledge of an AI system, and how that understanding changes when different feature-types occur in these systems. Both experiments showed that counterfactual explanations are effective in improving users' knowledge of an AI system's operation (as measured by predictive accuracy), more effective than causal explanations and the presentation of decisions without explanations. However, the subjective judgments of explanation satisfaction and trust did not align with users' task accuracy. Subjective judgments were not consistent across the two experiments, suggesting that such measurements may not be robust or reliable. Finally, both experiments showed that users understand decisions and explanations differently when they involve categorical or continuous features. In the following subsections, we discuss the implications of these findings with respect to several key issues: (i) the divergence between objective and subjective measures in XAI user studies (ii) the role of different feature-types – continuous versus categorical – in counterfactual algorithms, (iii) the future directions are suggested by these results.

6.1 Objective and Subjective Measures of Counterfactual and Causal Explanations

In both experiments, the counterfactual groups were more accurate than the control groups, and the accuracy of causal groups lay in between the other two. These findings suggest that counterfactuals help people reason about the causal importance of the features used in the system's decisions more effectively than mere re-descriptions of a decision, and slightly better than causal explanations. These effects emerge even when counterfactual and causal explanations contain matched information about how the query instance must change to cross the decision-boundary. This finding is consistent with evidence that counterfactuals elicit causal reasoning and enable people to understand causal relations [59, 62], and provides support for the use of counterfactuals in algorithmic recourse [26, 68], as they appear to aid user understanding of the predictions made by the system.

On the other hand, users' subjective explanation satisfaction and trust judgments did not correspond to their accuracy results. In Experiment 1, participants shown causal explanations gave the lowest judgments of satisfaction and trust across the groups, even though this group was more accurate than the control group. Surprisingly, in Experiment 2, the control groups, who received re-descriptions of the system's decisions and were least accurate in their predictions, reported the information they were provided with as satisfying and trustworthy, more so than groups who received explanations. This divergence may signal issues about the reliability of these particular satisfaction and trust scales (a point raised in other studies, see *e.g.*, [31]). However, the result also highlights the divergences between objective and subjective evaluations, already noted in other XAI user studies [4, 37, 65]. Indeed, Kuhl et al. [35, 36] have also recently reported positive effects of counterfactual explanations on task performance, but failure to find differences in subjective measures of helpfulness and usability in an abstract domain.

It is of some concern that participants who displayed the poorest understanding of the system's decisions (those in the control groups) often displayed the highest levels of satisfaction and trust, suggesting that shallow non-explanations may be enough to elicit users' trust and satisfaction without enabling them to learn about the system's operation. We suggest that our finding may be another manifestation of the *illusion of explanatory depth*, a phenomenon examined in research on the psychology of explanation that reveals a human tendency to overestimate one's causal understanding of complex phenomena[60]. The illusion can be dispelled by requiring people to generate their own explanation, or to answer diagnostic questions about the mechanism, or by providing them with detailed expert explanations. Hence a possible reason for the dissociation we have identified, that is, the lack of correspondence between understanding as measured by accuracy on the one hand, and subjective evaluation measures on the other hand, may be that this illusion of explanatory depth persists for those participants who receive no explanation, whereas users who are provided with explanations are prompted to integrate them with their existing mental model of the system and domain, thus enabling them to become aware of gaps in their understanding.

Overall, these results indicate that counterfactual explanations are a valuable tool to aid users in understanding automated systems and their decisions. Importantly, they also emphasise that where the objective of XAI is to improve human-machine team performance or achieve algorithmic recourse, as opposed to justifying an automated decision, it is crucial to probe user understanding of these decisions, rather than relying purely on users' self-reported evaluations.

6.2 Categorical Versus Continuous Features in Explanation

The current experiments indicate that users find features presented in a categorical format easier to understand and base predictions on than features presented in a continuous format. In Experiment 1, we found that users were more accurate in making predictions based on categorical features compared to continuous features. Experiment 2 demonstrated that showing people categorically-transformed features leads to higher accuracy (hence better understanding) than presenting features as continuous, indicating that this effect emerges from how features are presented rather than some inherent ontological property of the feature-type.

The findings have significant implications for counterfactual algorithms in XAI. Most XAI counterfactual methods (*e.g.*, [25, 55]) transform categorical features into continuous formats, using one-hot encoding or mapping to ordinal feature-spaces. These methods are commonly applied to tabular datasets, which have mixtures of continuous and categorical features, often involving high-stakes decision-making (*e.g.*, the COMPAS and German Credit datasets, for recidivism risk assessment and loan approval, respectively). The current results suggest that where possible, AI explanation should prioritise categorical features, to help users better understand decisions. However, the results also show that users can benefit from continuous features that have been transformed into categorical formats. Thus we expect that counterfactual methods that perform such transformations will be more effective in terms of user understanding compared to ones that do not. Recently we have developed one such method, which has been implemented and computationally evaluated, showing that it is possible to apply categorical transformations to counterfactual explanations, without significantly affecting explanatory coverage or effectiveness of the algorithm [69]. One consideration in this regard is that categorical features may impact users' subjective judgments; simplifying the features presented to users, may come at some cost of subjective trust in the system. Notably, the benefits of categorical features in explanation emerge regardless of explanation type, counterfactual or causal, and so this finding also has implications for XAI in general, and not solely for contrastive approaches.

6.3 Limitations and Future Directions

The present work emphasises how XAI can benefit from embracing evidence and methodologies from cognitive psychology. The findings detailed in this study suggest several possible future lines of research. First, the categorical features examined here were limited to binary values. Although these kinds of features occur in many datasets (*e.g.*, gender, Boolean values), categorical features can, in theory, have as many potential values as continuous ones. Moreover, depending on the system, features, user, or task context, a more fine-grained binning may be more appropriate than the binary division examined here. Hence, the current work presents a

clear baseline for the distinction between continuous and categorical features, but further research is necessary to establish whether there is a limit to the number of categorical values that humans can keep track of without compromising accuracy (that is, before the categorical features become as challenging as continuous ones). The difference in accuracy observed between the different features suggests that users may be able to monitor up to at least four categories (given that accuracy for *units* was higher than that for *weight* and *duration*), but future work should test this hypothesis. Second, when considering the kinds of features to alter in counterfactual explanations, methods have focused on properties such as mutability, actionability and plausibility, and users' understanding of continuous and categorical features has been assumed to be equal. One implication of our findings is the possibility that there may be other fundamental feature distinctions which have yet to be accounted for by XAI methods. Overall, the findings motivate a more psychologically-grounded and user-centric approach to XAI, to design methods that reflect the demonstrated benefits of counterfactual explanations and categorical features, as well as to evaluate the cognitive effects of explanations on users' understanding of AI systems.

ACKNOWLEDGMENTS

This research was supported by (i) the UCD Foundation, (ii) Science Foundation Ireland via the Insight SFI Research Centre for Data Analytics (12/RC/2289) and (iii) the Department of Agriculture, Food and Marine via the VistaMilk SFI Research Centre (16/RC/3835).

REFERENCES

[1] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 80–89. https://doi.org/10.1145/3351095.3372830

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020). https://doi.org/10.1016/j.inffus.2019.12.012

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[5] Bruce G. Buchanan and Edward H. Shortliffe. 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA.

[6] Ruth M.J. Byrne. 2005. *The Rational Imagination: How people create alternatives to reality.* MIT Press, Cambridge, MA.

[7] Ruth M.J. Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 6276–6282. https://doi.org/10.24963/ijcai.2019/876

[8] Ruth M.J. Byrne and Alessandra Tasso. 1999. Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory and Cognition* 27, 4 (1999), 726–740. https://doi.org/10.3758/BF03211565

[9] Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. 2019. Model agnostic contrastive explanations for structured data. *arXiv:1906.00117*. https://arxiv.org/abs/1906.00117

[10] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

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

[12] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. *arXiv*. http://arxiv.org/abs/2109.12480

[13] Orlando Espino and Ruth M.J. Byrne. 2018. Thinking About the Opposite of What Is Said: Counterfactual Conditionals and Symbolic or Alternate Simulations of Negation. *Cognitive Science* 42, 8 (2018), 2459–2501. https://doi.org/10.1111/cogs.12677

[14] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

[15] Tobias Gerstenberg, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum. 2021. A counterfactual simulation model of causal judgments for physical events. *Psychological Review* 128, 5 (2021), 936–975. https://doi.org/10.1037/rev0000281

[16] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2376–2384. https://proceedings.mlr.press/v97/goyal19a.html

[17] 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*. https://arxiv.org/abs/1805.10820

[18] Joseph Y. Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part I: Causes. 56, 4 (2005), 843–887. https://doi.org/10.1093/bjps/axi147

[19] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608*. http://arxiv.org/abs/1812.04608

[20] David Hume. 1748. *An Enquiry concerning Human Understanding* (a critical edition, 1999 ed.). Oxford University Press, Oxford, UK.

[21] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (2011), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

[22] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *arXiv:1907.09615*. http://arxiv.org/abs/1907.09615

[23] Daniel Kahneman and Dale T. Miller. 1986. Norm Theory. Comparing Reality to Its Alternatives. *Psychological Review* 93, 2 (1986), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

[24] Daniel Kahneman and Amos Tversky. 1982. The Simulation Heuristic. In *Judgment Under Uncertainty: Heuristics and Biases*, Daniel Kahneman, Paul Slovic, and Amos Tversky (Eds.). Cambridge University Press, New York, 201–8.

[25] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-Agnostic Counterfactual Explanations for Consequential Decisions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 895–905. https://proceedings.mlr.press/v108/karimi20a.html

[26] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse:contrastive explanations and consequential recommendations. *ACM Computing Surveys* 55, 5, Article 95 (Dec 2022), 29 pages. https://doi.org/10.1145/3527848

[27] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3442188.3445899

[28] 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. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4466–4474. https://doi.org/10.24963/ijcai.2021/609 Survey Track.

[29] Mark T. Keane and Barry Smyth. 2020. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In *Case-Based Reasoning Research and Development*, Ian Watson and Rosina Weber (Eds.). Springer International Publishing, Cham, 163–178. https://doi.org/10.1007/978-3-030-58342-2_11

[30] Frank C. Keil. 2006. Explanation and understanding. *Annual Review of Psychology* 57 (2006), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

[31] Eoin M. Kenny, Eoin D. Delaney, Derek Greene, and Mark T. Keane. 2021. Post-hoc Explanation Options for XAI in Deep Learning: The Insight Centre for Data Analytics Perspective. In *Pattern Recognition. ICPR International Workshops and Challenges*, Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (Eds.). Springer International Publishing, Cham, 20–34. https://doi.org/10.1007/978-3-030-68796-0_2

[32] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (2021), 103459. https://doi.org/10.1016/j.artint.2021.103459

[33] Eoin M. Kenny and Mark T Keane. 2021. On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11575–11585. https://doi.org/10.1609/aaai.v35i13.17377

[34] Lara Kirfel and Alice Liefgreen. 2021. What If (and How...)? - Actionability Shapes People's Perceptions of Counterfactual Explanations in Automated Decision-Making. In *ICML (International Conference on Machine Learning) Workshop on Algorithmic Recourse.* https://drive.google.com/file/d/1asi0PtgygYpJIAx2aiCG6OtldVvz7R2i/view

[35] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. In *FAccT 2022 - Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery (ACM), 2125–2137. https://doi.org/10.1145/3531146.3534630

[36] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Let's Go to the Alien Zoo: Introducing an Experimental Framework to Study Usability of Counterfactual Explanations for Machine Learning. *arXiv:2205.03398*. http://arxiv.org/abs/2205.03398

[37] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv:1902.00006*. https://arxiv.org/abs/1902.00006

[38] David A. Lagnado, Tobias Gerstenberg, and Ro'i Zultan. 2013. Causal responsibility and counterfactuals. *Cognitive Science* 37, 6 (2013), 1036–1073. https://doi.org/10.1111/cogs.12054

[39] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[40] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *arXiv:1707.01154*. https://arxiv.org/abs/1707.01154

[41] Matthew L. Leavitt and Ari Morcos. 2020. Towards falsifiable interpretability research. *arXiv:2010.12016* (2020). http://arxiv.org/abs/2010.12016

[42] David Lewis. 1973. Causation. *Journal of Philosophy* 70, 17 (1973), 556–567. https://doi.org/10.2307/2025310

[43] 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* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[44] Zachary C. Lipton. 2017. The Doctor Just Won't Accept That!. In *Interpretable ML Symposium, 31st Conference on Neural Information Processing Systems* (Long Beach, CA, USA). http://arxiv.org/abs/1711.08037

[45] Christopher G. Lucas and Charles Kemp. 2015. An improved probabilistic account of counterfactual reasoning. *Psychological Review* 122, 4 (2015), 700–734. https://doi.org/10.1037/a0039655

[46] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 90–98. https://doi.org/10.1145/3351095.3372824

[47] David R. Mandel and Darrin R. Lehman. 1996. Counterfactual Thinking and Ascriptions of Cause and Preventability. *Journal of Personality and Social Psychology* 71, 3 (1996), 450–463. https://doi.org/10.1037/0022-3514.71.3.450

[48] Keith D. Markman, Matthew N. McMullen, and Ronald A. Elizaga. 2008. Counterfactual thinking, persistence, and performance: A test of the Reflection and Evaluation Model. *Journal of Experimental Social Psychology* 44, 2 (2008), 421–428. https://doi.org/10.1016/j.jesp.2007.01.001

[49] David Martens and Foster Provost. 2014. Explaining data-driven document classifications. *MIS Quarterly: Management Information Systems* 38, 1 (2014), 73–99. https://doi.org/10.25300/MISQ/2014/38.1.04

[50] Rachel McCloy and Ruth M.J. Byrne. 2002. Semifactual "even if" thinking. *Thinking and Reasoning* 8, 1 (2002), 41–67. https://doi.org/10.1080/13546780143000125

[51] Alice McEleney and Ruth M.J. Byrne. 2006. Spontaneous counterfactual thoughts and causal explanations. *Thinking and Reasoning* 12, 2 (2006), 235–255. https://doi.org/10.1080/13546780500317897

[52] Aaron McGarvey. 2015. Easypower: sample size estimation for experimental designs. *R package version* 1, 1 (2015).

[53] Bjorn Meder, Tobias Gerstenberg, York Hagmayer, and Michael R. Waldmann. 2010. Observing and Intervening: Rational and Heuristic Models of Causal Decision Making. *The Open Psychology Journal* 3, 2 (2010), 119–135. https://doi.org/10.2174/1874350101003020119

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

[55] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. https://doi.org/10.1145/3351095.3372850

[56] Mike Oaksford and Keith Stenning. 1992. Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18, 4 (1992), 835. https://doi.org/10.1037/0278-7393.18.4.835

[57] Isabel Orenes, Orlando Espino, and Ruth M.J. Byrne. 2022. Similarities and differences in understanding negative and affirmative counterfactuals and causal assertions: Evidence from eye-tracking. *Quarterly Journal of Experimental Psychology* (2022). https://doi.org/10.1177/17470218211044085

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 13-17-August. 1135–1144. https://doi.org/10.1145/2939672.2939778

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

[60] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26, 5 (2002), 521–562. https://doi.org/10.1016/S0364-0213(02)00078-2

[61] Barry Smyth and Mark T. Keane. 2022. A Few Good Counterfactuals: Generating Interpretable, Plausible and Diverse Counterfactual Explanations. In *Case-Based Reasoning Research and Development*, Mark T. Keane and Nirmalie Wiratunga (Eds.). Springer International Publishing, Cham, 18–32. https://doi.org/10.1007/978-3-031-14923-8_2

[62] Barbara A. Spellman and David R. Mandel. 1999. When Possibility Informs Reality: Counterfactual Thinking as a Cue to Causality. *Current Directions in Psychological Science* 8, 4 (8 1999), 120–123. https://doi.org/10.1111/1467-8721.00028

[63] Girish H. Subramanian, John Nosek, Sankaran P. Raghunathan, and Santosh S. Kanitkar. 1992. A Comparison of the Decision Table and Tree. *Commun. ACM* 35, 1 (Jan 1992), 89–94. https://doi.org/10.1145/129617.129621

[64] Berk Ustun, Alexander Spangher, and Yang Liu. 2018. Actionable recourse in linear classification. In *Proceedings of the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning*,. 10–19. https://econcs.seas.harvard.edu/files/econcs/files/spangher_fatml18.pdf

[65] 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 (2021), 103404. https://doi.org/10.1016/j.artint.2020.103404

[66] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable Counterfactual Explanations Guided by Prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano (Eds.). Springer International Publishing, Cham, 650–665. https://doi.org/10.1007/978-3-030-86520-7_40

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

[68] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, April 2018 (2018). https://doi.org/10.2139/ssrn.3063289

[69] Greta Warren, Barry Smyth, and Mark T. Keane. 2022. "Better" Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI). In *Case-Based Reasoning Research and Development*, Mark T. Keane and Nirmalie Wiratunga (Eds.). Springer International Publishing, Cham, 63–78. https://doi.org/10.1007/978-3-031-14923-8_5

[70] Erik Matteo Prochet Widmark. 1981. *Principles and applications of medicolegal alcohol determination.* Biomedical Publications, Davis, CA.

A SUBJECTIVE JUDGMENT SCALES

A.1 Explanation Satisfaction Scale

Table 1: Explanation Satisfaction Scale

Explanation Satisfaction Scale Items
1. From the explanation, I understand how the app works.
2. The explanation of how the app works is satisfying.
3. The explanation of how the app works has sufficient detail.
4. This explanation of how the app works seems complete.
5. This explanation of how the app tells me how to use it.
6. This explanation of how the app works is useful to my goals.
7. This explanation of the app shows me how accurate the app is.
8. This explanation lets me judge when I should trust and not trust the app.

A.2 Trust Scale

Table 2: Trust Scale

Trust Scale Items
1. I am confident in the app. I feel that it works well.
2. The outputs of the app are very predictable.
3. The app is very reliable. I can count on it to be correct all the time.
4. I feel safe that when I rely on the app I will get the right answers.
5. The app is efficient in that it works very quickly.
6. I am wary of the app. (reverse scored)
7. The app can perform the task better than a novice human user.
8. I like using the system for decision making.

B MATERIALS

B.1 Lower and upper quartile values for continuous features

Table 3: Lower and upper quartile values for continuous features

Feature	Lower quartile value	Upper quartile value
Units	4	5
Weight	69kg	83kg
Duration	79 mins	106 mins

B.2 Example explanations: Experiment 2

Table 4: Examples of counterfactual explanations in Experiment 2

Feature	Mixed features	Categorical features
Units	If Richard had drunk 3 units more, he would have been over the limit.	If Richard had drunk more units, he would have been over the limit.
Weight	If Samantha's weight had been 50kg lighter, she would have been over the limit.	If Samantha's weight had been lighter, she would have been over the limit.
Gender	If Jenny's gender had been male, she would have been under the limit.	If Jenny's gender had been male, she would have been under the limit.
Stomach-fullness	If Kevin's stomach had been fuller, he would have been under the limit.	If Kevin's stomach had been fuller, he would have been under the limit.

Table 5: Examples of causal explanations in Experiment 2

Feature	Mixed features	Categorical features
Units	Because Richard drank 3 units too few, he was under the limit.	Because Richard drank too few units, he was under the limit.
Weight	Because Samantha's weight was 50kg too heavy, she was under the limit.	Because Samantha's weight was too heavy, she was under the limit.
Gender	Because Jenny's gender was female, she was over the limit.	Because Jenny's gender was female, she was over the limit.
Stomach-fullness	Because Kevin's stomach was too empty, he was over the limit.	Because Kevin's stomach was too empty, he was over the limit.

Table 6: Examples of control descriptions in Experiment 2

Feature	Mixed features	Categorical features
Units	Richard was under the limit.	Richard was under the limit.
Weight	Samantha was under the limit.	Samantha was under the limit.
Gender	Jenny was over the limit.	Jenny was over the limit.
Stomach-fullness	Kevin was over the limit.	Kevin was over the limit.