Features of Explainability: How Users Understand Counterfactual and Causal Explanations for Categorical and Continuous Features in XAI

Greta Warren [1,2], Mark T. Keane[1,2,3] and Ruth M.J. Byrne [4]

[1] *School of Computer Science, University College Dublin, Dublin, Ireland*
[2] *Insight SFI Centre for Data Analytics, University College Dublin, Dublin, Ireland*
[3] *VistaMilk SFI Research Centre, University College Dublin, Dublin, Ireland*
[4] *School of Psychology and Institute of Neuroscience, Trinity College Dublin, University of Dublin, Dublin, Ireland*

Abstract

Research on eXplainable AI (XAI) has recently focused on the use of counterfactual explanations to address interpretability, recourse, and bias in AI decisions. Many proponents of these counterfactual algorithms claim they are cognitively valid in their generation of "plausible" explanations using "important", "actionable" or "causal" features, where these features are computed from the model being explained. However, very few of these claims have been tested by psychological studies; specifically, claims about the role of different feature-types have not been validated, perhaps suggesting that a more considered analysis of these knowledge representations is required. In this paper, we consider the cognitive validity of a key representational distinction, between *continuous* and *categorical features*, in counterfactual explanations. In a controlled user study (N=127), we tested the effects of counterfactual and causal explanations on the *objective* accuracy of users' predictions of the decisions made by a simple AI system, and their *subjective* judgments of satisfaction and trust in the explanations. We found that users understand explanations referring to categorical features more readily than those referring to continuous features. We also discovered a dissociation between objective and subjective measures: counterfactual explanations elicit higher accuracy of predictions than no-explanation control descriptions but no higher accuracy than causal explanations, and yet counterfactual explanations elicit greater satisfaction and trust judgments than causal explanations. We discuss the implications of these findings for cognitive aspects of knowledge representation in XAI.

Keywords
XAI, counterfactual explanation, algorithmic recourse, interpretable machine learning

1. Introduction

The use of automated decision making in computer programs that impact people's everyday lives has led to rising concerns about the fairness, transparency, and trustworthiness of Artificial Intelligence (AI) [1,2]. These concerns have created renewed interest in, and an urgency about, tackling the problem of eXplainable AI (XAI), that is, the need to provide explanations of AI systems' decisions. Recently, counterfactual explanations have been advanced as a promising solution to the XAI problem because of their compliance with data protection regulations, such as the EU's General Data Protection Regulation (GDPR) [3], their potential to support algorithmic recourse [4], and their psychological importance in explanation [5,6]. The prototypical XAI scenario for counterfactuals is the explanation of an automated decision when a bank customer's loan application is refused; on querying the decision,

IJCAI-ECAI'22 Workshop: Cognitive Aspects of Knowledge Representation, July 23–29, 2022, Vienna, Austria
EMAIL: greta.warren@ucdconnect.ie (G. Warren); mark.keane@ucd.ie (M. T. Keane); rmbyrne@tcd.ie (R. M. J. Byrne)
ORCID: 0000-0002-3804-2287 (G. Warren); 0000-0001-7630-9598 (M. T. Keane); 0000-0003-2240-1211 (R. M. J. Byrne)

CEUR Workshop Proceedings (CEUR-WS.org)

the customer is told "*if you had asked for a lower loan of $10,000, your application would have been approved*". These counterfactual explanations appear to be readily understood by humans, while also offering users possible recourse to change the decision's outcome (e.g., by lowering their loan request). Although there is now a substantial XAI literature on counterfactuals, because of a lack of user studies we know very little about *how* people understand these counterfactual explanations of AI decisions, and *which* aspects of counterfactual methods are critical to their use in XAI. Many counterfactual algorithms aim to explain decisions by referring to "plausible", "actionable", or "causally important" features, however, it is unclear how to reliably identify these sorts of features, much less how, and which (if any) of these characteristics are important to users. In this paper, we focus on the representational distinction between *continuous* and *categorical* features in a statistically well-powered and psychologically well-controlled study (N=127), examining how different explanations impact people's understanding of automated decisions. We test explanations of automated decisions about blood alcohol content and legal limits for driving using counterfactual explanations (e.g., "*if John had drunk 3 units instead of 5 units, he would have been under the limit*"), compared to causal explanations (e.g., "*John was over the limit because he drank 5 units*"), and descriptions ("*John was over the limit*"). The study examines not only the effects of explanations but also the effects of different types of features – categorical features (gender, stomach-fullness) and continuous features (units, duration of drinking, body weight). It includes objective measures of the accuracy of participants' understanding of the automated decision, and subjective measures of their satisfaction and trust in the system and its decisions. In the remainder of this introduction, we consider the relevant related work in this area on counterfactual explanations in XAI (see 1.1), as well as how feature-types (see 1.2), and causal explanations (see 1.3) have been handled in these systems, before outlining the current experiment (see 1.4).

1.1. Counterfactual Explanations

In recent years, XAI research on the use of counterfactuals has exploded, with over 100 distinct computational methods proposed in the literature (for reviews see [7,8]). These various techniques argue for different approaches to counterfactual generation; some advance optimisation techniques, [3,9] others emphasise the use of causal models [10], distributional analyses [11] or the importance of instances [12]. These alternative proposals are typically motivated by claims that the method in question generates "good" counterfactuals for end-users; for instance, that the counterfactuals are psychologically "good" because they are proximal [3], plausible/actionable [10], sparse [12] or diverse [9]. However, most of these claims are based on intuition rather than on empirical evidence. A recent review found that just 21% of 117 papers on counterfactual explanation included any user-testing, and fewer (only ~7%) tested specific properties of the method proposed [7]. This state of affairs raises the possibility that many of these techniques contain functions with little or no psychological validity, that may have no practical benefit to people in real-life applications [13].

Consider what we *have* learned from the few user studies on counterfactuals in XAI. Most user studies test whether counterfactual explanations impact people's responses relative to no-explanation controls or some other explanation strategy (e.g., example-based explanations or rule-based explanations [14]). These studies assess explanation quality using "objective" measures (e.g., user predictive accuracy) and/or "subjective" measures (e.g., user judgments of trust, satisfaction, preference). In philosophical and psychological research, explanations are understood to be designed to change people's understanding of the world, events or phenomena [5,15]. In XAI, this definition has been conceptualised to mean that explanation should improve people's understanding of the AI system, the domain involved in the task and/or their performance on the target task [16]. An explanation is effective, therefore, if people *objectively* perform better on a task involving the AI system by, for example, being faster, more accurate, or by being able to predict what the system might do next [14,17–20]. Concretely, if a person with diabetes is using an application to estimate their blood sugar levels for insulin treatments, ideally the system's predictions would help them better understand their condition in the future; for example, their predictions of their own blood sugar levels should improve, when the application's help is not available.

So far, a handful of studies have shown mixed support for the use of counterfactual explanations in improving user understanding in this regard. "*What-if*" counterfactual explanations have been found to improve performance in prediction and diagnosis tasks relative to no-explanation controls, however they did not improve performance appreciably more than other explanation options (“*why-not”*, *“how-to”* and “*why”* explanations) [17]. Visual counterfactual explanations were shown to increase classification accuracy relative to no-explanation controls in a small sample of users [18]. In some cases, prompting users to reason counterfactually about a decision may impair objective performance. One study compared counterfactual tasks, in which users were asked if a system’s recommendation would change given a perturbation of some input feature, to simulation tasks, in which users were asked to predict the recommendation based on the input features [19]. Counterfactuals 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 [20]. These findings are consistent with the proposal that counterfactuals require people to consider multiple possibilities, to compare reality to the suggested alternative and to infer causal relations, as is often reported in the cognitive psychological literature [21]. They also provide further evidence that counterfactuals aid people to reason about past decisions, and prepare for future ones, but require cognitive effort and resources [22-25]. However, some caution should be exercised in generalising from the small collection of XAI user studies on counterfactual explanations, given the diversity of tasks, domains and experimental designs; some do not involve controls, and many others use too few test items or very small numbers of participants to be confident about the findings reported.

XAI research has also focused on whether explanations work *subjectively*; that is, whether the explanation improves people’s trust or satisfaction in the AI system, or whether the explanation makes people “feel better” about their interaction with the system, with generally positive results. Users judge counterfactuals as more appropriate and fair than example-based [2], demographic-based, and influence-based explanations [1]. Providing contrastive explanations in a sales-forecasting domain has also been found to increase self-reported understanding of the system’s decisions [20]. However, two studies have shown dissociations between objective and subjective measures in XAI. Users shown contrastive rule-based explanations self-reported better understanding of the system’s decision than no-explanation controls, however neither of these groups, nor users shown contrastive example-based explanations, showed any improvement in accuracy for predicting what the system might do, and tended to follow the system’s advice, even when incorrect [14]. A similar disconnect between objective and subjective evaluation measures was found for tasks that systematically increased the complexity of a system’s causal rules; although users’ response times and judgments of difficulty also increased, little effect of complexity was observed on task accuracy [19]. Thus, in XAI, studies asking users how well they understand a system’s decisions or how satisfying they find an explanation, may not accurately reflect the true explanatory power of different sorts of explanations, particularly given people’s propensity to overestimate their understanding of complex causal mechanisms [26]. Notably, if an explanation strategy has no objective impact on understanding but is subjectively preferred by users, then concerns about its ethical use could arise. In the current study, we assess the extent to which counterfactual and causal explanations increase people’s understanding, using measures that are objective (accuracy in predicting what the system will do) and subjective (i.e., judgments of trust and satisfaction).

1.2. Feature-Types in Explanations

Advocates of counterfactual explanation methods often emphasise the role of different feature-types in making explanations “good” or “psychologically plausible”. Many counterfactual methods distinguish between the types of features to be used in explanations; arguing that it makes sense to use features that are *mutable* rather than *immutable* [27] (e.g., being told to “*reduce your age to get a loan*” is not useful). Furthermore, proponents argue that the features used in the counterfactual should be *causally important* [8] and/or *actionable* [10]; a counterfactual explanation proposing to reduce the size of the requested loan is more actionable and therefore better than one telling the customer to modify a long-standing, bad credit-rating. However, to our knowledge, there is only one existing study that examines users’ assessments of feature-types in XAI [28], which found that while the mutability/actionability of a feature is predictive of user satisfaction and (self-reported) understanding, the importance of this factor varies depending on the domain. Although such feature-distinctions are made readily in AI models, from a cognitive perspective they appear context-dependent and ill-defined [29,30]. For example, the different sorts of mutability are more ambiguous than assumed by computational approaches. Indeed, psychologically, perhaps there are more fundamental representational distinctions to be made between feature properties, such as whether people can understand continuous features (such as income or credit-score) or categorical features (such as race or gender) equally well. Psychological studies have long shown that people do not tend to spontaneously make changes to continuous variables such as time or speed, e.g., when they imagine how an accident could have been avoided [31]. This representational distinction between continuous and categorical features is important, because if people are less likely to manipulate continuous features or have difficulties understanding counterfactuals about them, then the potential causal importance or actionability of such features is moot. For example, in algorithmic recourse, people may better understand counterfactual advice that says, "*you need to change your credit-score from bad to good*" rather than advice that says "*you need to increase your credit score from 3.4 to 4.6*". To date, the differential impacts of categorical and continuous feature-types has not been considered in counterfactual methods for XAI. Most counterfactual generation methods make the assumption that users treat and understand continuous and categorical features in the same way (e.g., DiCE [9] applies one-hot encoding to categorical feature values). In the present study, we examine people's understanding of counterfactual explanations for different feature-types (continuous versus categorical), predicting that explanations focusing on categorical features will be more readily understood, leading to greater predictive accuracy based on these features.

1.3. Causal Explanations

A third goal of the present work is to compare counterfactual explanations to those using causal rules (with respect to feature-types), as the latter are a long-standing explanation strategy in AI. In philosophical and psychological research, there is consensus that everyday explanations often invoke some notion of cause and effect [15]. Causal explanations and counterfactuals have long been viewed as being intertwined in complex ways [27,32,33], although psychologically they differ in significant ways. For example, when people create causal explanations, they focus on causes that are sufficient and may be necessary for an outcome to occur, whereas when they create counterfactuals they focus on causes that are necessary but may not be sufficient [23]. In AI, causal explanations are often cast as IF-THEN rules (e.g., in expert systems such as MYCIN [34] or decision trees [19, 35]). In XAI, it is commonly claimed that such rule-based explanations are inherently interpretable, although some have pointed out that this claim may not be accurate [13]. One study reports that when users were given causal decision sets for a system, they achieved high accuracy in a prediction task, however in a counterfactual task using the same decision sets, participants' accuracy decreased, while their response times and judgments of subjective difficulty increased [19]. Another study found that contrastive rule-based explanations were effective in helping users identify the crucial feature in a system's decision and increased people's sense of understanding the system; indeed contrastive rule-based explanations were more effective than contrastive example-based explanations [14]. These findings suggest that causal rules may be as good as, if not sometimes better than, counterfactual explanations in some task contexts. In the cognitive psychological literature, it has been found that when people are asked to reflect on an imagined negative event, they spontaneously generate twice as many causal explanations as counterfactual thoughts about it [36], consistent with the proposal that causal explanations may not require people to compare multiple possibilities in the way that counterfactual explanations do [37,38]. Since it has also been shown that people make difficult inferences from counterfactuals more readily than they do from their factual counterparts [39], it is plausible that counterfactuals' evocation of multiple possibilities may help users consider an AI system's decision more deeply. Given the clear importance of these two explanation options – causal and counterfactual strategies – both are compared to one another in the present study. Considering that their appeal to the purportedly contrastive nature of explanation [5] is one of the main arguments for the use of counterfactuals in XAI, and given the psychological evidence that counterfactuals are understood by thinking about more possibilities than causal explanations, we predict that counterfactual explanations will aid users in understanding the system's decisions more than causal explanations, and that both will outperform mere descriptions of an outcome.

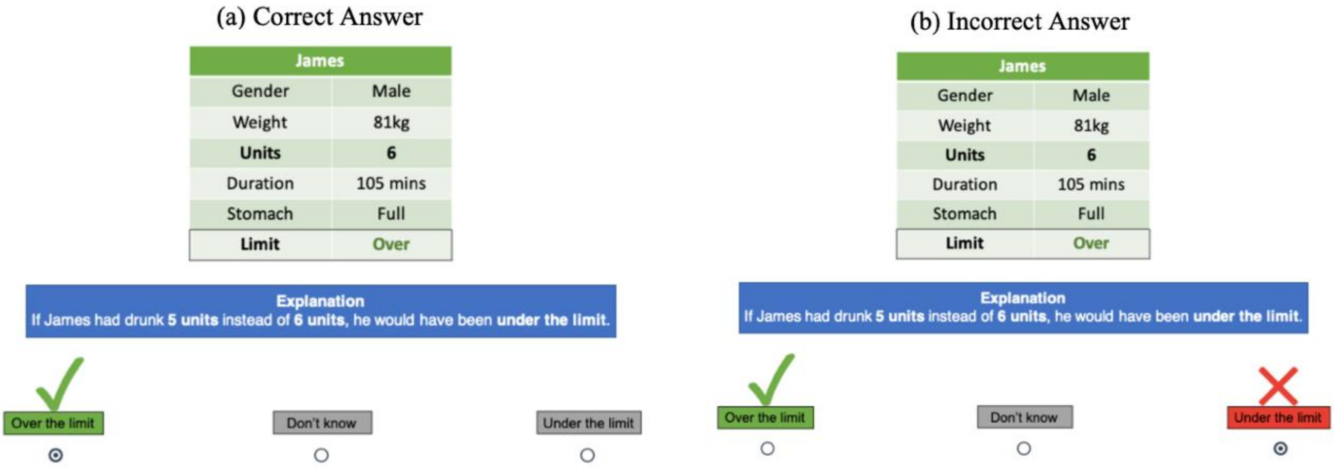
1.4. Outline of Current Study

The study tested the impact of counterfactual versus causal explanations, and continuous versus categorical features, on users' accuracy of understanding and subjective evaluation of a simulated AI system designed to predict blood alcohol content and legal limits. Participants were shown predictions by the system for different instances, with explanations (e.g. "*If Mary had weighed 80kg instead of 75kg, she would have been under the limit.*"). The study consisted of two phases (i) a *training phase* in which participants were asked to predict the system's decision (i.e. an individual being over or under the legal blood alcohol content threshold to drive a car), and were provided with feedback on the system's predictions and with explanations for each decision, and (ii) a *testing phase,* in which they were asked to predict outcomes for a different set of test instances, this time with no feedback nor any explanations. In the training phase, participants considered the system's predictions and learned about the blood alcohol content domain with the help of the explanations, to determine whether this experience *objectively* improved their understanding of the domain. The testing phase objectively measured their developed understanding of the system by measuring the accuracy of their predictions. Users' subjective evaluations were also recorded by measuring their judgments of satisfaction and trust.

2. The Task: Predicting Legal Limits for Driving

Participants were presented with the output of a simulated AI system presented as an application, designed to predict whether someone is over the legal blood alcohol content limit to drive. The system relies on a commonly-used approximate method, the Widmark equation [40], that uses five key features for blood alcohol content with the limit threshold being 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, from which the study's materials were drawn (N=2000).

In the experimental task, participants were instructed that they would be testing a new application, *SafeLimit,* designed to inform people whether or not they are over the legal limit to drive, from 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. In the training phase, participants were shown examples of tabular data for different individuals, 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. The order of these options was randomised, to ensure that participants did not merely click on the same button-order each time. After giving their response, feedback was given on the next page, with the correct answer highlighted using a green tick-mark, and the incorrect answer (if selected) highlighted using a red X-mark (see Figures 1 and 2). Above the answer options, participants were also shown an explanation, and which explanation they were shown depended on the experimental condition. Figures 1 and 2 show sample materials used in the counterfactual and causal conditions, respectively. Note that in both conditions the explanations draw attention to a key feature (e.g., the units drunk) as being critical to the prediction made. In all the study's conditions, a balanced set of instances were used, with eight items for each of the five features presented. Upon completing the training phase, participants began the testing phase (see Figure 3). Again, they were shown instances referring to individuals (different to those in the training phase), and asked to judge if the individual was over or under the legal limit to drive. After submitting their 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; for instance, "*Given this person's WEIGHT, please make a judgment about their blood alcohol level*." Again, in this phase, a balanced set of instances was used, with eight items for each of the five features presented.


(a) Correct Answer

James	
Gender	Male
Weight	81kg
Units	**6**
Duration	105 mins
Stomach	Full
Limit	**Over**

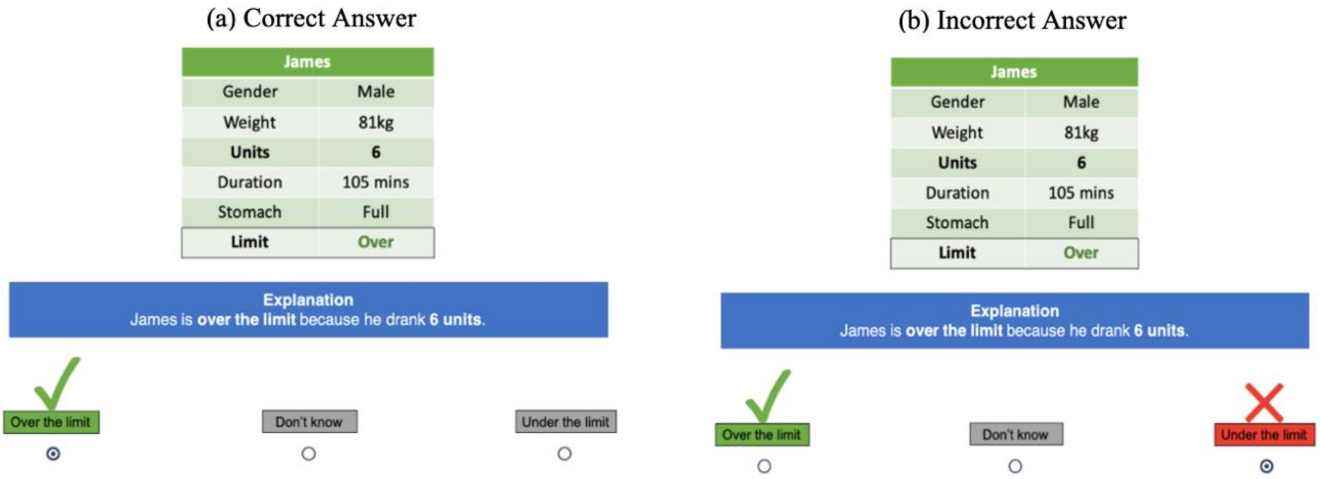
Explanation
If James had drunk **5 units** instead of **6 units**, he would have been **under the limit**.

(b) Incorrect Answer

James	
Gender	Male
Weight	81kg
Units	**6**
Duration	105 mins
Stomach	Full
Limit	**Over**

Explanation
If James had drunk **5 units** instead of **6 units**, he would have been **under the limit**.

Figure 1: Feedback for (a) Correct Answer and (b) Incorrect Answer in the Counterfactual condition.


(a) Correct Answer

James	
Gender	Male
Weight	81kg
Units	**6**
Duration	105 mins
Stomach	Full
Limit	**Over**

Explanation
James is **over the limit** because he drank **6 units**.

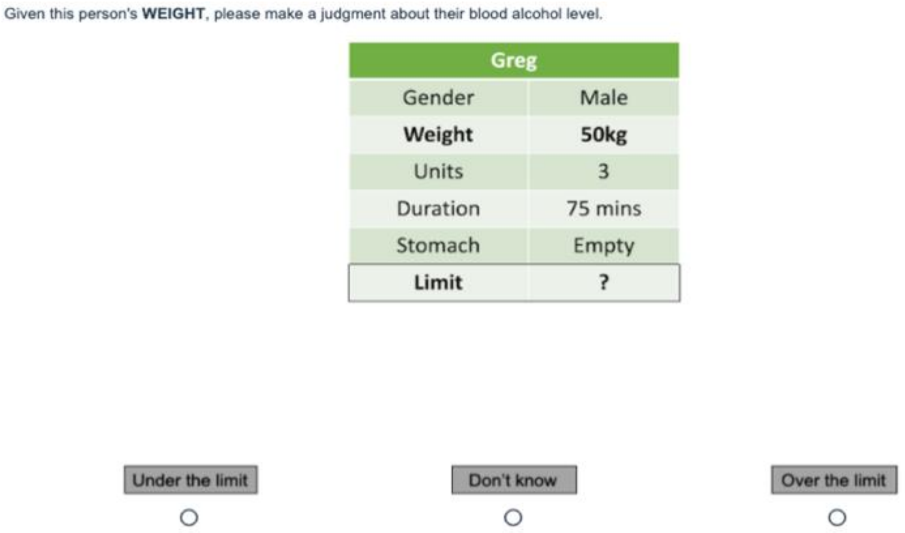
(b) Incorrect Answer

James	
Gender	Male
Weight	81kg
Units	**6**
Duration	105 mins
Stomach	Full
Limit	**Over**

Explanation
James is **over the limit** because he drank **6 units**.

Figure 2: Feedback for (a) Correct Answer and (b) Incorrect Answer in the Causal condition.

The objective measure of performance in both phases of the study was accuracy (i.e., correct predictions made by participants compared to those of the system). The subjective measures were explanation satisfaction and trust in the system, assessed using the DARPA project's Explanation Satisfaction and Trust scales [16] respectively). To assess engagement with the task, participants completed four attention checks at random intervals throughout the experiment, and were asked to recall the 5 features used by the application by selecting them from a list of 10 options at the end of the session.



Given this person's **WEIGHT**, please make a judgment about their blood alcohol level.

Greg	
Gender	Male
Weight	**50kg**
Units	3
Duration	75 mins
Stomach	Empty
Limit	**?**

Figure 3: Example of a prediction task in the testing phase.

2.1. Method

We compared the impact of counterfactual and causal explanations, to descriptions of the system's decisions as a control condition, on the predictions people made about the *SafeLimit* application's decisions. Participants were assigned in fixed order to one of three groups (counterfactual, causal, control) and completed the experiment, consisting of (i) a *training phase* in which they made predictions and were given feedback with explanations or descriptions and (ii) a *testing phase* where they made predictions with no feedback and no explanations (for all groups). Hence, any observed differences in accuracy in the testing phase should reflect people's understanding of the AI system based on their experiences in the training phase, which differed only in the nature of the explanation (or control description) provided. Participants were presented with 40 items in each phase, which were 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 training phase will be more accurate than performance in the testing phase, (ii) counterfactual explanations will improve accuracy more than causal explanations, as they are potentially more informative, (iii) predictions about categorical features will be more accurate than predictions about continuous features, if people find the former less complex than the latter, and (iv) counterfactual explanations will be judged as more satisfying and trustworthy than causal explanations, given previous studies showing that they are often subjectively preferred over other explanations.

2.1.1. Participants and Design

The participants (N=127), crowdsourced using the Prolific platform (https://www.prolific.co/), were randomly assigned to the three between-participant conditions: counterfactual explanation (n=41), causal explanation (n=43) and control (n=43). These groups consisted of 80 women, 46 men, and one non-binary person aged 18-74 years (*M*=33.54, *SD*=13.15); and were pre-screened to select native English speakers from Ireland, the United Kingdom, the United States, Australia, Canada and New Zealand, who had not participated in previous related studies. 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. A further 11 participants were excluded prior to any data analysis, one for giving identical responses for each trial, and 10 who failed more than one attention or memory check. Before testing, the power analysis with G*Power [41] indicated that 126 participants were required to achieve 90% power for a medium-sized effect with alpha <.05 for two-tailed tests. Ethics approval for the study was granted by the University College Dublin ethics committee with the reference code LS-E-20-11-Warren-Keane.

2.1.2. Materials and Procedure

Eighty instances were randomly selected, based on key filters from the 2000-item dataset generated for the blood alcohol content domain (based on stepped increments of a feature's normally-distributed values with realistic upper/lower limits). Specifically, the procedure randomly selected an instance (query case) and incrementally increased or decreased one of the five feature's values until its blood alcohol content value crossed the decision boundary to create a counterfactual case. For the categorical features, *gender* and *stomach-fullness*, the inverse value was assigned, while continuous variables were incremented in steps of 15kg for *weight*, 15 minutes for *duration* and 1 *unit* for alcohol. If the query case could not be perturbed to cross the decision boundary, a different case was randomly selected, and the procedure was re-started. 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*", with the *control group* given a description of the outcome (e.g., "*John is over the limit*"). This selection procedure was performed 16 times for each feature, a total of 80 times, with the further constraint that an equal number of instances were found on either side of the decision boundary (i.e., equal numbers under and over the limit). Each instance was then randomly assigned to one of two sets of materials, each comprising 40 items, again ensuring an equal number of instances were classified as under/over the limit. To avoid any material-specific confounds, the 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 read detailed instructions about the tasks (available at https://osf.io/j7rm3/) and completed one practice trial for each phase of the study before commencing. They then progressed through the presented instances, randomly re-ordered for each participant, within the training and testing phases. After completing both phases, they completed the Explanation Satisfaction and Trust scales. Participants were debriefed and paid £2.61 for their time. The experiment took approximately 28 minutes to complete.

2.2. Results and Discussion

The results show that providing explanations improved the accuracy of people's predictions, and that categorical features led to higher prediction accuracy than continuous features. Participants' accuracy on categorical features was markedly higher in the testing phase than 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). Participants judged counterfactual explanations to be more satisfying and trustworthy than causal explanations, however counterfactual explanations had only a slightly greater impact than causal explanations on participants' accuracy in predicting the AI system's decisions. The data for this experiment are publicly available at https://osf.io/wqdtn/.

2.2.1. Analysis of the Accuracy Measure

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 on the proportion of correct answers given by each participant (see Figure 4). A Huynh-Feldt correction was applied to the main effect of Feature and its interactions. Significant main effects were found for Explanation, $F(2,124)=5.63$, $p=.005$, $\eta_p^2=.083$, for Task, $F(1,124)=32.349$, $p<.001$, $\eta_p^2=.207$, and for Feature, $F(3.945, 489.156)=47.599$, $p<.001$, $\eta_p^2=.277$. Task interacted with

Feature, $F(4, 496)=7.23$, $p<.001$, $\eta_p^2=.055$. No other effects were significant[1]. These effects were further examined in post hoc analyses.

First, with respect to the main effect of Explanation, 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. However, the Causal group (*M*=.614, *SD*=.09) did not differ significantly from the Counterfactual, *p*=.245, or Control groups, *p*=.186. Further exploratory analysis indicated there was a reliable trend in increasing accuracy with the following ordering of the groups for their accuracy scores (Page's L(40)=1005.0, *p*<.001): Counterfactual > Causal > Control. These results suggest that providing explanations is better than not providing them, for improving accuracy. They also show, as predicted, that counterfactual explanations have a greater impact than causal explanations, and compared to a control condition given no explanations. Note that these effects were observed for both phases of the study overall (Explanation does not interact with Task).

Second, with respect to the significant Task and Feature main effects, and their significant interaction, the decomposition of the interaction revealed that accuracy improves 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 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, all other features did not differ significantly from each other (*p*>.05 for all comparisons). In the testing phase, similar tests found accuracy to be 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. Furthermore, accuracy for *stomach-fullness* was significantly higher than that 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.[2]

Further exploratory analysis indicates it is the diversity in the range of feature values that may lead to these 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 this rank-ordering 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.

[1] No other two-way interactions were reliable, neither Explanation with Task, *F*(2, 124)=.759, *p*=.47, nor Explanation with Feature, *F*(7.89, 489.156)=1.14, *p*=.335, nor was the three-way interaction significant, *F*(8, 496)=1.215, *p*=.288.

[2] 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.

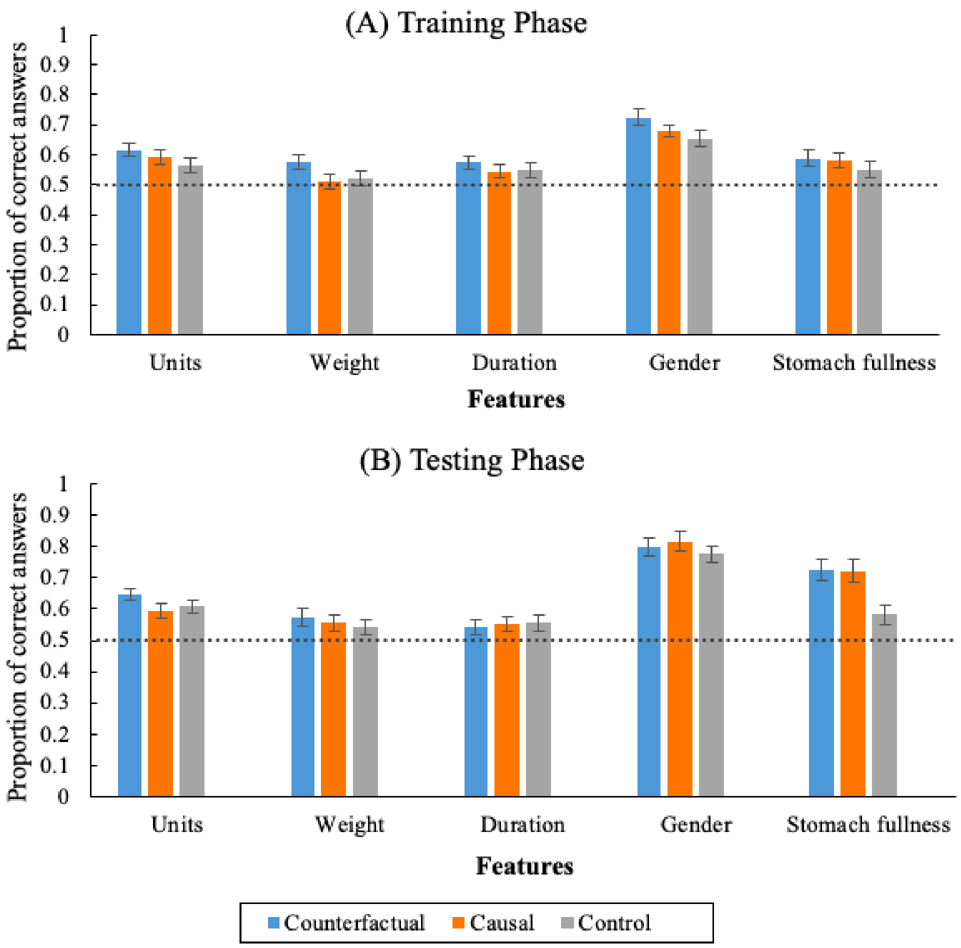


Figure 4: Mean accuracy (proportion of correct answers) across conditions for each feature in the (A) Training and (B) Testing phases of the study. Error bars are standard error of the mean; dashed line represents chance accuracy.

2.2.2. Analysis of the Subjective Measures: Satisfaction and Trust

All groups completed the DARPA Explanation Satisfaction and Trust scales after completing the two main phases in the experiment (see Figure 5).

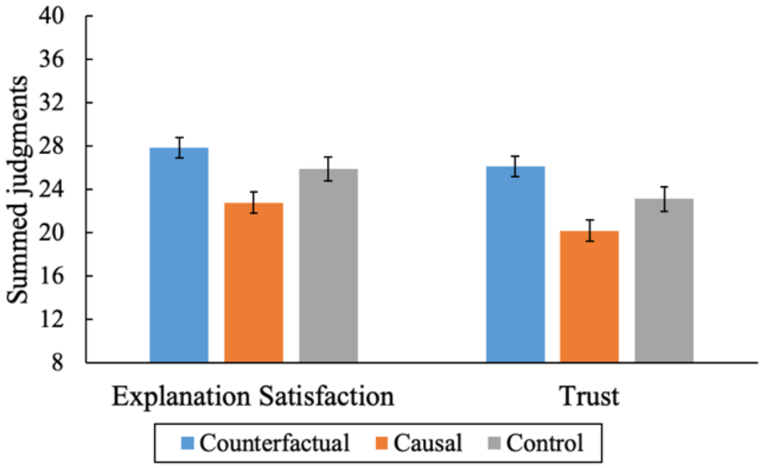


Figure 5: Summed judgments for Explanation Satisfaction and Trust scales. Error bars are standard error of the mean.

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. Significant differences between the three groups were identified $F(2, 126)=6.104$, $p=.003$, $\eta_p^2=.09$. Post hoc Tukey HSD tests showed that the counterfactual group ($M=27.83$, $SD=6.12$) gave significantly higher satisfaction judgments than 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. 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 descriptions, and descriptions were slightly more satisfying than causal explanations. People were less satisfied with causal explanations compared to counterfactual explanations or even none at all.

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. Significant differences between the groups were identified $F(2, 126)=8.184$, $p<.001$, $\eta_p^2=.117$. Post hoc Tukey HSD tests showed that the counterfactual group ($M=26.15$, $SD=6.14$) gave significantly higher trust judgments than 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 ($p=.115$). A reliable trend was identified when rank-ordering judgments for each item in the order: Counterfactual > Control > Causal, L(8)=112.0, $p<.001$. Similar to the satisfaction judgments, these results suggest that counterfactual explanations were somewhat more trustworthy than descriptions, and descriptions were slightly more trustworthy than causal explanations. People placed less trust in causal explanations compared to counterfactual explanations or even none at all .

3. General Discussion, Conclusions and Future Directions

The present study shows that a knowledge representation distinction between abstract feature-types – continuous versus categorical – is cognitively significant in impacting people's understanding of explanations in XAI, a distinction whose significance is not noted in current counterfactual methods. The experiment showed that users' accuracy in predicting a system's decisions improved when they were provided with explanations compared to none at all, and when they were provided with counterfactual explanations compared to causal ones; counterfactual explanations were also subjectively preferred compared to causal ones. The experiment also shows that users' accuracy in predicting a system's decisions improved when they relied on categorical features rather than continuous features, with improvements over time between the training and test phases of the study. In the following sub-sections, we discuss the implications of these findings for (i) the significance of categorical and continuous features in explanations, and (ii) the role of explanations in XAI and the relative differences between counterfactual and causal explanations (and descriptions).

3.1. The Primacy of Categorical Over Continuous Features

The results described in this paper indicate that users were more accurate in making predictions based on categorical features than continuous features within each phase of the experiment. User accuracy increased in the testing phase relative to the training one, but this rise was mainly due to improvement in making predictions about categorical features (*gender* and *stomach-fullness*), an improvement that does not occur for continuous features (*units, duration, weight*). We cannot attribute this effect to the provision of explanations (the three-way interaction was not significant); instead it is an improvement that emerges as people gain more experience throughout the training phase with the categorical features. Current counterfactual methods in XAI do not recognise any functional benefits for categorical features over continuous ones. These counterfactual methods transform categorical features to allow them to be processed similarly to continuous ones, using one-hot encoding or by mapping to ordinal feature spaces. Hence, no current model recognises that one feature-type might be more psychologically beneficial than another. Remarkably, given the 100+ methods in the counterfactual XAI literature, no current algorithm gives primacy to categorical features over continuous ones for explanations of the predictions of an AI system. Many models consider mutability and actionability as being important to the provided counterfactual explanations but neither of these concepts account for the results found here. Recall, the results showed improved performance for the *gender* and *stomach-fullness* features even though the former is immutable and non-actionable (in the context of blood alcohol decisions) and the latter is mutable and actionable. Moreover, the results showed less improved performance for the *units, duration,* and *weight* features, even though they are mutable and actionable (albeit *weight* in the context of blood alcohol decisions is immutable in the short term). Hence, the improvement in accuracy for the *gender* and *stomach-fullness* features over the course of the experiment (from training to testing phase) is more plausibly due to their simplicity (both have just two possible feature values) compared to the more complex continuous features (which have many possible feature values). There are clear implications of these results for counterfactual approaches in algorithmic recourse; namely, that it would be better to focus on categorical features than on continuous ones when the predictive outcomes are equivalent.

3.2. What Is It That (Counterfactual) Explanations Do?

The results also have a bearing on cognitive aspects of explanations. There is an increasing recognition that explanations can play one of several roles in XAI. One major role is to improve the user's understanding of the domain, the AI system, or both, manifested by objective performance improvements in the task domain when explanations are provided. Measuring effects of explanation on objective performance is the guiding proposal in Hoffman's et al.'s [16] conceptual framework for XAI and is a repeated theme in XAI user studies [13,16]. However, a number of studies indicate that explanations, especially counterfactual explanations, may *not* improve objective performance on the task [14,19]. Moreover, many studies show that explanations are more likely to impact subjective assessments than objective performance; that is, people tend to self-report higher understanding [20] or judge decisions to be more fair [1] or appropriate [2]. These considerations combined with the findings of the present study, raise potentially serious ethical concerns about the use of explanations. They suggest that some explanations may cause people to "feel better" about the AI system, without gaining any insight into why it made a prediction or how it works. Explanations may lead the recipient to a somewhat false assessment of the value of the system, akin to the "illusion of explanatory depth", wherein people overestimate their understanding of causal mechanisms underlying common phenomena [26], potentially leading to inappropriate trust in a system and its decisions. The present results help to clarify the role that explanations may take. Overall, the counterfactual group were more accurate than the control group, and the causal group's accuracy lay in-between the other groups. This observation suggests that counterfactuals help people reason about the causal importance of the features used in the system's decisions more effectively than mere descriptions of an outcome, and slightly better than causal explanations. Moreover, counterfactual explanations improved people's accuracy in both phases, without depending on transfer or learning from the training to the testing phase (i.e., there was an effect of Explanation, but this factor did not interact with any other factor). This conclusion is highly consistent with key findings in the psychological literature that counterfactuals elicit causal reasoning and enable people to understand causal relations [24,32]. Indeed, these findings also support proposals for the use of counterfactuals in algorithmic recourse [3,8], as they seem to better prompt an understanding of the predictions made by the system.

3.3. Future Directions

The present work emphasises how AI needs to consider the cognitive aspects of knowledge representation; it shows that a cognitively-blind AI will miss functional aspects of proposed algorithms that have major cognitive effects. Several issues need to be addressed further in future work. First, the categorical features examined here were limited to binary values. Although these kinds of features commonly occur in many datasets (such as gender, ethnicity, or Boolean true/false features), categorical features can, in theory, have as many potential values as continuous ones. Hence, it 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 features). The differences in accuracy observed between the different types of 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 further investigation is needed to test this hypothesis. A further question is whether people find categorical features easier to reason about because of feature-value diversity or some other property of categorical features. Overall, the findings motivate a more psychologically-grounded approach to counterfactuals in XAI, to design methods that reflect the demonstrated cognitive benefits of categorical features, based on experimentally corroborated hypotheses rather than on untested conjectures.

4. Acknowledgements

This paper resulted from research funded by (i) the UCD Foundation and (ii) Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics (12/RC/2289-P2).

5. References

[1] J. Dodge, Q. Vera Liao, Y. Zhang, R. Bellamy, C. Dugan, Explaining models: An empirical study of how explanations impact fairness judgment, in: Int. Conf. Intell. User Interfaces, Proc. IUI, volume Part F1476, 2019, p. 275–85. doi:10.1145/3301275.3302310.

[2] R. Binns, M. Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, It’s reducing a human being to a percentage, Proc (2018) 1–14. doi:10.1145/3173574.3173951.

[3] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv J Law Technol 31 (2018).

[4] A. H. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: From counterfactual explanations to interventions, in: FAccT 2021 - Proc 2021 ACM Conf Fairness, Accountability, Transpar, 2021.

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

[6] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, IJCAI Int. Jt. Conf. Artif. Intell (2019) 6276–82. doi:10.24963/ijcai. 2019/876.

[7] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques, in: IJCAI-21, 2021.

[8] A. H. Karimi, G. Barthe, B. Schölkopf, I. Valera, A survey of algorithmic recourse: contrastive explanations and consequential recommendations, ACM Comput Surv 1 (2021) 1–26.

[9] R. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, FAT* 2020 (2020) 607–17. doi:10.1145/3351095.3372850.

[10] A. H. Karimi, G. Barthe, B. Balle, I. Valera, Model-agnostic counterfactual explanations for consequential decisions, Proc 108 (2020).

[11] E. M. Kenny, M. T. Keane. On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning, AAAI-21, 11575–11585.

[12] M. T. Keane, S. B. Counterfactuals, W. Find Them, Int, Conf. Case-Based Reason., Springer, Cham, 2020.

[13] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017.

[14] J. der Waa, E. Nieuwburg, A. Cremers, N. M. XAI, A comparison of rule-based and example-based explanations, Artif Intell 291 (2021). doi:10.1016/j.artint.2020.103404.

[15] F. C. Keil, Explanation and understanding, Annu Rev Psychol 57 (2006) 227–54. doi:10. 1146/annurev.psych.57.102904.190100.

[16] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, 2018.

[17] B. Y. Lim, A. K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, Conf. Hum. Factors Comput. Syst. - Proc (2009) 2119–28. doi:10.1145/1518701.1519023.

[18] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, Int. Conf. Mach. Learn (2019) 2376–84.

[19] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, Human evaluation of models built for interpretability, Proc. AAAI Conf. Hum. Comput. Crowdsourcing (2019) 59–67.

[20] A. Lucic, H. Haned, M. Rijke, Why does my model fail? Contrastive local explanations for retail forecasting, in: FAT*2020, 2020 pp. 90–8. doi:10.1145/3351095.3372824.

[21] R. M. J. Byrne, Counterfactual thought, Annu Rev Psychol 67 (2016) 135–57. doi:10.1146/annurev-psych-122414-033249.

[22] R. M. J. Byrne, The Rational Imagination, MIT Press, Cambridge, MA, 2005.

[23] D. R. Mandel, D. R. Lehman, Counterfactual thinking and ascriptions of cause and preventability, J Pers Soc Psychol 71 (1996) 450–63. doi:10.1037/0022-3514.71.3.450.

[24] N. J. Roese, K. Epstude, The functional theory of counterfactual thinking: New evidence, new challenges, new insights. In Advances in experimental social psychology (2017) 56, pp. 1-79. AP.

[25] K. D. Markman, M. N. McMullen, R. A. Elizaga, Counterfactual thinking, persistence, and performance: A test of the reflection and evaluation model. Journal of Experimental Social Psychology (2008), 44(2), 421-428.

[26] L. Rozenblit, F. C. Keil, The misunderstood limits of folk science: An illusion of explanatory depth, Cogn Sci 26 (2002) 521–62. doi:10.1016/S0364-0213(02)00078-2.

[27] D. Kahneman, D. T. Miller, Norm theory: Comparing reality to its alternatives, Psychol Rev 93 (1986) 136–53. doi:10.1037/0033-295X.93.2.136.

[28] L. Kirfel, Liefgreen. A. What if (and how...)? Actionability shapes people's perceptions of counterfactual explanations in automated decision-making, in: ICML (International Conf, Learn. Work. Algorithmic Recourse, Mach, 2021.

[29] V. Girotto, D. Ferrante, S. Pighin, M. Gonzalez, Postdecisional counterfactual thinking by actors and readers. Psychological Science 18(6) (2007) 510-515.

[30] S. Pighin, R. M. Byrne, D. Ferrante, M. Gonzalez, V. Girotto, Counterfactual thoughts about experienced, observed, and narrated events. Thinking & Re. 17(2) (2011). 197-211.

[31] D. Kahneman, A. Tversky, The simulation heuristic, in: D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, CUP, New York, 1982, pp. 201–8.

[32] Spellman, B. A., & Mandel, D. R. (1999). When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions in Psychological Science*, *8*(4), 120-123.

[33] J. Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach, Part I: Causes. Br J Sci 56 (2005) 843–87. doi:10.1093/bjps/axi147.

[34] B. G. Buchanan, E. H. Shortliffe, Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project, CUMINCAD, 1984.

[35] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, DecisSupport Syst 51 (2011) 141–54. doi:10.1016/j.dss.2010.12.003.

[36] A. McEleney, R. M. J. Byrne, Spontaneous counterfactual thoughts and causal explanations, Think Reason 12 (2006) 235–55. doi:10.1080/13546780500317897.

[37] D. A. Lagnado, T. Gerstenberg, R. I. Zultan, Causal responsibility and counterfactuals. Cognitive science 37(6) (2013) 1036-1073.

[38] C. R. Walsh, R. M. Byrne, How people think "if only…" about reasons for actions. Thinking & Reasoning, 13(4), (2007) 461-483.

[39] R. M. J. Byrne, A. Tasso, Deductive reasoning with factual, possible, and counterfactual conditionals, Mem Cogn 27 (1999) 726–40. doi:10.3758/BF03211565.

[40] E. M. P. Widmark, Die theoretischen Grundlagen und die praktische Verwendbarkeit der gerichtlich-medizinischen Alkoholbestimmung, Urban Schwarzenberg, Berlin, 1932.

[41] F. Franz, E. Erdfelder, A. Buchner, A. Lang, Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses, Behavior research methods 41 (4) (2009) 1149-1160.