Explaining Model Confidence Using Counterfactuals

Thao Le, Tim Miller, Ronal Singh, Liz Sonenberg

School of Computing and Information Systems, The University of Melbourne
thaol4@student.unimelb.edu.au, {tmiller, rr.singh, l.sonenberg}@unimelb.edu.au

Abstract

Displaying confidence scores in human-AI interaction has been shown to help build trust between humans and AI systems. However, most existing research uses only the confidence score as a form of communication. As confidence scores are just another model output, users may want to understand why the algorithm is confident to determine whether to accept the confidence score. In this paper, we show that counterfactual explanations of confidence scores help study participants to better understand and better trust a machine learning model's prediction. We present two methods for understanding model confidence using counterfactual explanation: (1) based on counterfactual examples; and (2) based on visualisation of the counterfactual space. Both increase understanding and trust for study participants over a baseline of no explanation, but qualitative results show that they are used quite differently, leading to recommendations of when to use each one and directions of designing better explanations.

Introduction

Explaining why an AI model gives a certain prediction can promote trust and understanding for users, especially for non-expert users. While recent research (Zhang, Liao, and Bellamy 2020; Wang, Zhang, and Lim 2021) has used confidence (or uncertainty) measures as a way to improve AI model understanding and trust, the area of explaining why the AI model is confident (or not confident) in its prediction is still underexplored (Tomsett et al. 2020).

In Machine Learning (ML), the *confidence score* indicates the chances that the ML model's prediction is correct. In other words, it shows how *certain* the model is in its prediction, which can be defined as the predicted probability for the best outcome (Zhang, Liao, and Bellamy 2020). Another way to define the *confidence score* is based on uncertainty measures, which can be calculated using entropy (Bhatt et al. 2021) or using *uncertainty sampling* (Lewis and Gale 1994), (Monarch 2021, p93).

In this paper, we complement prior research by applying a counterfactual (CF) explanation method to generate explanations of the confidence of a predicted output. It is increasingly accepted that explainability techniques should be built on research in philosophy, psychology and cognitive science (Miller 2019; Byrne 2019) and that the evaluation process of explanation should involve human-subject studies (Miller, Howe, and Sonenberg 2017; Förster et al. 2021; Kenny et al. 2021; van der Waa et al. 2021). We therefore evaluate our explanation to know whether counterfactual explanations can improve *understanding*, *trust*, and *user satisfaction* in two user studies using existing methods for assessing understanding, trust and satisfaction. We present the CF explanation using two designs: (1) providing counterfactual examples (example-based counterfactuals); and (2) visualising the counterfactual space for each feature and its effect on model confidence (visualisation-based counterfactuals).

Our contributions are:

- We formalise two approaches for the counterfactual explanation of confidence score: one using counterfactual examples and one visualising the counterfactual space.
- Through two user studies we demonstrate that showing counterfactual explanations of confidence scores can help users better understand and trust the model.
- Using qualitative analysis, we observe limits of the two explainability approaches and suggest directions for improving presentations of counterfactual explanations.

Background and Related Work

In this section, we review related work on counterfactual explanations and confidence (or uncertainty) measures.

Counterfactual Explanations

Counterfactual explanation is described as the possible smallest changes in input values in order to change the model prediction to a desired output (Wachter, Mittelstadt, and Russell 2017). It has been increasingly used in explainable AI (XAI) to facilitate human interaction with the AI model (Miller 2019, 2021; Byrne 2019; Förster et al. 2021). Counterfactual explanations can be expressed in the following example: "You were denied a loan because your annual income was $30,000. If your income had been $45,000, you would have been offered a loan". To generate counterfactuals, (Wachter, Mittelstadt, and Russell 2017) suggest finding solutions of the following loss function.

$$\arg\min_{x'} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') \tag{1}$$

where x' is the counterfactual solution; $(f(x') - y')^2$ presents the distance between the model's prediction output of counterfactual input x' and the desired counterfactual output y'; $d(x, x')$ is the distance between the original input and the counterfactual input; and λ is a weight parameter. A high λ means we prefer to find counterfactual point x' that gives output $f(x')$ close to the desired output y', a low λ means we aim to find counterfactual input x' that is close to the original input x even when the counterfactual output $f(x')$ can be far away from the desired output y'. In this model, $f(x)$ would be the predicted output, such as a denied loan, and y' would be the desired output – the loan is granted. The counterfactual x' would be the properties of a similar customer that would have received the loan. Equation 1 can be solved by using the Lagrangian approach. However, this approach has stability issues (Russell 2019). Therefore, Russell (2019) proposes another search algorithm to generate counterfactual explanations based on mixed-integer programming, assumed where input variables can be continuous or categorical values. They defined a set of linear integer constraints, which is called *mixed polytope*. These constraints can be given to Gurobi Optimization (Gurobi Optimization, LLC 2023) and then an optimal solution is generated.

Antorán et al. (2021) propose Counterfactual Latent Uncertainty Explanations (CLUE), to identify features responsible for the model's uncertainty. Their idea for showing counterfactual examples is similar to ours, however we go further by considering ways to visualise the counterfactual space, run a more comprehensive user study to measure understanding, satisfaction, and trust, and undertake a qualitative analysis to identify limitations of current approaches.

There are many other approaches to solving counterfactuals for tabular (Mothilal, Sharma, and Tan 2020; Keane and Smyth 2020), image (Goyal et al. 2019; Dhurandhar et al. 2018), text (Jacovi et al. 2021; Riveiro and Thill 2021) and time series data (Delaney, Greene, and Keane 2021a). None of these are for explaining model confidence, however, the underlying algorithms could be modified to search over the model confidence instead of the model output.

Confidence (Uncertainty) Measures

A confidence score measures how confident a ML model is in its prediction; or inversely, how uncertain it is. A common method of measuring uncertainty is to use the prediction probability (Delaney, Greene, and Keane 2021b; Bhatt et al. 2021). Specifically, *uncertainty sampling* (Lewis and Gale 1994) is an approach that queries unlabelled instance x with maximum uncertainty to get human feedback. There are four types of uncertainty sampling (Monarch 2021, p70): *Least confidence*, *Margin of confidence*, *Ratio of confidence* and *Entropy*. Zhang, Liao, and Bellamy (2020) demonstrate that communicating confidence scores can support trust calibration for end users. Wang, Zhang, and Lim (2021) also argue that showing feature attribution uncertainty helps improve model understanding and trust.

van der Waa et al. (2020) propose a framework called *Interpretable Confidence Measures (ICM)* which provides predictable and explainable confidence measures based on case-based reasoning (Atkeson, Moore, and Schaal 1997). Case-based reasoning provides prediction based on similar past cases of the current instance. This approach however did not address counterfactual explanations of model confidence.

Formalising Counterfactual Explanation of Confidence

This section describes two methods for CF explanation: one based on counterfactual examples (Antorán et al. 2021) and one based on counterfactual visualisation as in Figure 1.

Generating Counterfactual Explanation of Confidence

In this section, we show how to generate counterfactual explanations of the confidence score in data where input variables can take either categorical or continuous values. The counterfactual model can generate explanations to either increase or decrease the confidence score of a specific class. For example, when the AI model predicts that an employee will leave the company with confidence of 70%, a person may ask: *Why is the model 70% confident instead of 40% confident or less?*. This person could ask why the model prediction did not have a lower confidence score when they were sceptical about the high confidence score. We aim to generate counterfactual inputs that bring the confidence score to 40% or lower. An example of counterfactual explanation in this case is: *"One way you could have got a confidence score of 40% instead is if Daily Rate had taken the value 400 rather than 300"*. Therefore, from this counterfactual explanation, we know that we can achieve lowering of the confidence of them resigning from the company by increasing the employee's daily rate.

We now describe our approach to generate counterfactuals for confidence scores. We follow Russell (2019) in proposing an algorithm to search for counterfactual points of output confidence. Importantly, we modify this approach to find counterfactual points that change the confidence score but do not change the predicted class.

Formally, given a question: "Why does the model prediction have a confidence score of $U(x)$ rather than greater than (or less than) T?" where T is a user-defined confidence threshold, x is the input instance, $U(x)$ is the confidence score of the original prediction, we want to find the counterfactual explanation of confidence $U(x')$ generated by data point x' such that $U(x') > T$ or $U(x') < T$ depending on the question. In case the user cannot give a threshold T, the default threshold T value is the original confidence score $U(x)$ of the prediction. We seek the counterfactual point x' by solving Equation 2:

$$\arg\min_{x'}(||x - x'||_{1,w} + |U(x') - T|) \tag{2}$$

such that:

$$U(x') > T \qquad \text{if } T > U(x) \tag{3}$$

$$U(x') < T \qquad \text{if } T < U(x) \tag{4}$$

$$\begin{cases} P(y = k \mid x') < D & \text{if } P(y = k \mid x) < D \\ P(y = k \mid x') \geq D & \text{if } P(y = k \mid x) \geq D \end{cases} \tag{5}$$

Attribute	Alternative 1	Alternative 2	Original
Marital status	-	-	Married
Years of education	-	-	9
Occupation	**Manager**	**Skilled Specialty**	**Service**
Age	-	-	63
Any capital gains	-	-	No
Working hours per week	-	-	12
Education	-	-	High School
Confidence score	**30.1%**	**42.1%**	**57.8%**
AI prediction	**Lower than $50,000**		

Table 1: Example-based counterfactual explanation presented in a table. In alternative columns, notation (-) means the value is unchanged from the original value, we only highlight the values that changed.

where $||.||_{1,w}$ is a weighted l_1 norm with weight w defined as the inverse median absolute deviation (MAD) (Wachter, Mittelstadt, and Russell 2017); D is the decision boundary that classifies the class.

We apply Equation 3 when we want to find counterfactual x' that increases the confidence score, and Equation 4 for a counterfactual x' that decreases the confidence score. Since x and x' will give the same output prediction as class k but different confidence scores $U(x)$ and $U(x')$, $P(x)$ and $P(x')$ must be in the same space according to the decision boundary, defined as Equation 5.

Example-Based Counterfactual Explanation

Given the original instance input shown in column *Original Value* in Table 1, the AI model predicts that this person has an income of *Lower than* $\$50,000$ with a confidence score of 57.8%. We choose a factual confidence score $T = 45\%$ and search for x' where $U(x') < T$. An example of counterfactual explanation generated using our method is: *"One way you could have got a confidence score of less than* 45% *(*30.1%*) instead is if Occupation had taken value Manager rather than Service."*

We presented counterfactuals in a table, such as in Table 1. We show the details of a person in column *Original Value* and the prediction that their income is lower than $\$50,000$. When we change the value of feature *Occupation* as in columns *Alternative 1* and *Alternative 2*, the confidence score changes but the prediction is still lower than $\$50,000$. From this table, we can find the correlation between the *Occupation* and the confidence score; the occupation *Service* gives the prediction with the highest confidence score among all three occupations.

Visualisation-Based Counterfactual Explanation

In this section, we propose a method for visualising the counterfactual space of a model and how this affects the model's confidence as shown in Figure 1 and 2. The idea is to visualise how varying a single feature affects the model's confidence, relative to the factual input x. For example, Figure 1 shows the visualisation based on Table 1 in the income

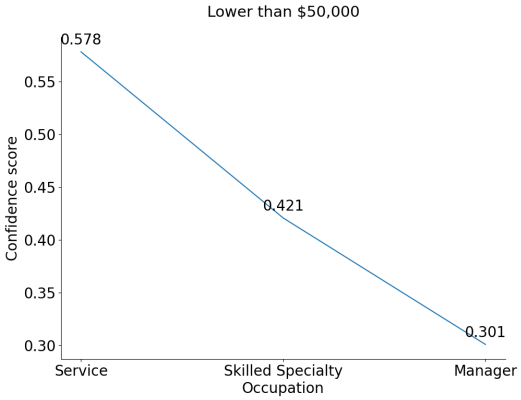

Figure 1: Counterfactual visualisation: Categorical variable

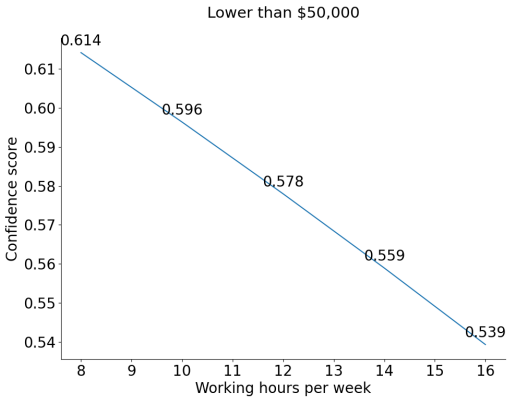

Figure 2: Counterfactual visualisation: Continuous variable

prediction task. Here we can see the prediction reaches maximum confidence score at *Service* occupation. The title of this graph shows the output prediction *Lower than* $\$50,000$ and the feature name *Occupation* which we used to change the values.

This visualisation technique is based on the idea of **Individual Conditional Expectation (ICE)** (Goldstein et al. 2015). ICE is often used to show the effect of a feature value on the predicted probability of an instance. In our study, we show how changing a feature value can change the *confidence score* instead of changing the predicted probability as in the original ICE. There are two types of variables in the dataset: (1) categorical variable, and (2) continuous variable. So we define the ICE for confidence score of a single feature x_i of instance x such that $F(x_i) = U(x_i)$ for all x_i, where:

- $x_i \in D$ if x_i is a categorical value and D is the categorical set
- $x_i \in [c_{\min}, c_{\min}+t, \ldots, c_{\max}]$ if x_i is a continuous value; $c_{\min}$ and $c_{\max}$ are the minimum and maximum values of a continuous range and t is a fixed increment.

If we use only a 2-dimensional graph, we can visualisation the changes of only one feature, whereas counterfactual examples can explain how changing multiple features simultaneously affect the confidence. However, visualising

	Control (C)	Treatment - Example-Based (E)	Treatment - Visualisation-Based (V)
Phase 1	Participants are given plain language statement, consent form and demographic questions (age, gender)		
Phase 2	Participants are provided with		
	Input instances	Input instances	Input instances
	AI model's prediction class	AI model's prediction class	AI model's prediction class
		Counterfactual examples	Counterfactual visualisation
Phase 3	Nothing	10-point Likert *Explanation Satisfaction Scale*	
Phase 4	10-point Likert *Trust Scale*		

Table 2: Summary of participants' tasks in our three experimental conditions

the counterfactual space allows us to easily identify the lowest and highest confidence values for categorical values and the trend of continuous values.

Human-Subject Experiments

Our user experiments test the following hypotheses.

- **Hypothesis 1a/b (H1a/b)**: **Example-based/Visualisation-based** counterfactual explanations help users better **understand** the AI model than when they are not given explanations.
- **Hypothesis 2a/b (H2a/b)**: **Example-based/Visualisation-based** counterfactual explanations help users better **trust** the AI model than when they are not given explanations.

It is necessary to test against the baseline of no explanation because providing explanations is not always useful compared to not providing any explanations (Lai and Tan 2019; Bansal et al. 2021). We then evaluate the difference between example-based counterfactual explanations and visualisation-based counterfactual explanations based on the following hypotheses.

- **Hypothesis 3a/b/c (H3a/b/c)**: *Visualisation-based* counterfactual explanations help users better **understand/trust/be satisfied with** the AI model than *example-based* counterfactual explanations.

To evaluate **understanding**, i.e., H1a, H1b and H3a, we use *task prediction* (Hoffman et al. 2018, p11). Participants are given some instances and their task is to decide for which instance the AI model will predict a higher confidence score. Thus, task prediction helps evaluate the user's mental model about their understanding in model confidence.

To evaluate **trust**, i.e., H2a, H2b and H3b, we use the 10-point Likert *Trust Scale* from (Hoffman et al. 2018, p49). For **satisfaction**, i.e., H3c, we use the 10-point Likert *Explanation Satisfaction Scale* from (Hoffman et al. 2018, p39).

Experimental Design

Dataset We ran the experiment on two different domains from two different datasets, which are *income prediction domain* and *HR domain*. Both datasets are selected so that experiments can be conducted on general participants with no requirement of particular expertise. The data used for the income prediction task is the Adult Dataset published in UCI Machine Learning Repository (Dua and Graff 2017) that includes 32561 instances and 14 features. This dataset classifies a person's income into two classes (below or above \$50K) based on personal information such as marital status, age, and education. In the second domain, we use the IBM HR Analytics Employee Attrition Performance dataset published in Kaggle (Pavansubhash 2017), which includes 1470 instances and 34 features. This dataset classifies employee attrition as yes or no based on some demographic information (job role, daily rate, age, etc.). We selected the seven most important features for both datasets by applying the Gradient Boosting Classification model over all data.

Attribute	Employee 1	Employee 2	Employee 3
Marital status	Married	Married	Married
Years of education	15	15	15
Occupation	**Service**	**Manager**	**Skilled Specialty**
Age	25	25	25
Any capital gains	No	No	No
Working hours per week	30	30	30
Education	Bachelors	Bachelors	Bachelors
AI model prediction	**Lower than \$50,000**		

Table 3: Example input instances provided in the question. The question is: "For which employee the AI model predicts with the highest confidence score?"

Model Implementation In our experiments, we use logistic regression to calculate the probability of a class, so $P(x) = \frac{1}{1+e^{-y}}$ where $y = wx$ is a linear function of point x. We chose logistic regression because of its simplicity so that we can easily define the confidence score. Moreover, although logistic regression models are considered intrinsically interpretable models (Molnar 2019), it is still challenging to reason about their behaviour when we want to have a lower (or higher) confidence score. In future work, our studies can be extended to using counterfactual tools for more complex models, such as CLUE (Antorán et al. 2021).

We choose *margin of confidence*, which is the difference between the first and the second highest probabilities (Monarch 2021, p93) as the formula of confidence score $U(x)$. The higher the difference between two class probabilities, the more confident the prediction is in the highest probability class.

Procedure Before conducting the experiments, we received ethics approval from our institution. We recruited participants on Amazon Mechanical Turk (Amazon MTurk), a

popular crowd-sourcing platform for human-subject experiments (Buhrmester, Kwang, and Gosling 2016). The experiment was designed as a Qualtrics survey[1] and participants can navigate to the survey through the Amazon MTurk interface. We allowed participants 30 minutes to finish the experiment and paid each participant a minimum of USD $7 for their time, plus a maximum of up to USD $2 depending on their final performance.

We use a between-subject design such that participants were randomly assigned into one of three groups: (1) *Control (C)*; (2) *Treatment with Example-Based Explanation (E)*; or (3) *Treatment with Visualisation-Based Explanation (V)*. For each group, there are four phases that are described in Table 2. The difference between the control group and the treatment group is that in the control group, participants were not given any explanations. In the task prediction (phase 2), participants in the control group were only shown input values along with the AI model prediction class as in Table 3. In the treatment group, participants were provided with either example-based explanations (e.g. Table 1) or visualisation-based explanations (e.g. Figure 1). The participants each received the same 10 questions. For each question, they were asked to select an input instance out of *3 instances* that the AI model would predict with the highest confidence score. A question can have either one or two explanations depending on the number of modified attributes in the question. For instance, the question in Table 3 changes only one attribute *Occupation* so participants were given a single explanation in treatment conditions. An explanation can either present a *categorical variable* (e.g. Figure 1) or a *continuous variable* (e.g. Figure 2).

We scored each participant using: 1 for a correct answer, -2 for a wrong answer and 0 for selecting "I don't have enough information to decide". To imitate high-stake domains, the loss for a wrong choice is higher than the reward for a correct choice (Bansal et al. 2019, p2433). They are also asked to briefly explain why they choose that option in a text box, which is analysed later in the qualitative analysis. The final compensation was calculated based on the final score — a score of 0 or less than 0 received $7 USD and no bonus. A score greater than 0 received a bonus of $0.2 for each additional score.

Participants We recruited a total of 180 participants for two domains, that is 90 participants for each domain from Amazon MTurk. Then 90 participants were evenly randomly allocated into three groups (30 participants in each group). All participants were from the United States. We only recruited Masters workers, who achieved a high degree of success in their performance across a large number of Requesters [2]. For the *income prediction domain*, 41 participants were women, 1 was self-specified as non-binary, 48 were men. Between them, 4 participants were between Age 18 and 29, 34 between Age 30 and 39, 27 between Age 40 and 49, 25 over Age 50. For the *HR domain*, 43 participants were women, 47 were men. Age wise, 4 participant was between Age 18 and 19, 37 between Age 30 and 39, 26 between Age 40 and 49, 23 over Age 50.

[1]https://www.qualtrics.com/

[2]https://www.mturk.com/worker/help

	Understanding			Trust			Satisfaction
	H1a E	**H1b** V	**H3a** E vs V	**H2a** E	**H2b** V	**H3b** E vs V	**H3c** E vs V
Domain 1	✓	✓	×	✓	✓	✓	✓
Domain 2	✓	✓	×	✓	✓	×	×

Table 4: Summary of hypothesis tests in two domains. ✓ represents the hypothesis is supported, × represents the hypothesis is rejected.

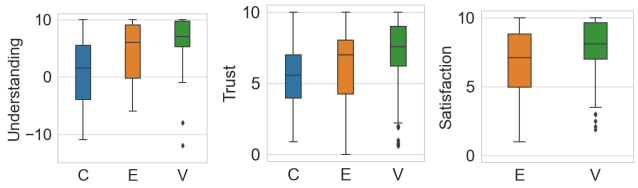

Figure 3: Domain 1 (Income)

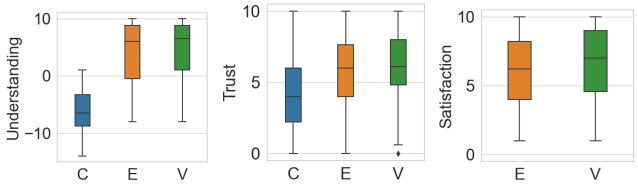

Figure 4: Domain 2 (HR)

We performed *power analysis* for two independent sample t-test to determine the needed sample sizes. We calculate the Cohen's d between control and treatment group and obtain the effect size of 0.7 and 0.67 in income and HR domain. Using power of 0.8 and significant alpha of 0.05, we get sample sizes of 26 and 29 in the two domains. Thus, we determine the sample size needed for a group is 30 and the total number of samples needed is 90 for one domain.

Results: Summary of Both Domains

In this section, we present the results from our experiment for two domains that used the income and HR datasets. We tested for data normality by using the Shapiro-Wilks test and found that our data was not normally distributed. Therefore, we applied the Mann–Whitney U test, which is a non-parametric test equivalent to the independent samples t-test to perform pairwise comparisons between two groups. Table 4 summarises our results of testing the seven hypotheses. Figure 3 and 4 show the results of the two studies.

The results show that counterfactual explanations of confidence scores help users understand and trust the AI model more than those who were not given counterfactual explanations. We conclude that **H1a, H1b, H2a** and **H2b** are supported in both studies ($p < 0.005, r > 0.41$).

There is no statistically significant difference in im-

Code	Definition
W-Reversed category (CAT)	The participant selected the instance that has a lowest confidence score instead of a highest confidence score among all instances
W-Linear assumption (CAT)	Assumed the correlation between confidence score and attribute values was linear when it was not (e.g. the feature was categorical)
W-Small differences (CAT & CON)	Selected a wrong answer due to small differences in the explanation and/or the question
W-Reversed correlation (CON)	Reversed the trend of the explanation of a continuous variable
W-Case-based (CON)	Used case-based reasoning when the correlation was linear
D-No correlation (CAT & CON)	Could not find the trend of the confidence score
D-Different attribute values (CAT & CON)	Argued that the values of instances in the explanations are not the same as values in the question
D-Outside range (CON)	The modified values in the question are beyond the lowest and highest values in the explanation
C-Correlation-based (CON)	Found the correlation in the explanation
C-Case-based (CAT)	Got the correct answer based on examples in the explanation without mentioning about the correlation

Table 5: The codebook for participants' responses to evaluate how they understand the provided explanations. *CAT*, *CON* mean the code is applied for categorical variables and continuous variables, respectively. *W* corresponds to wrong answers. *D* corresponds to the "do not have enough information to decide". *C* corresponds to correct answers.

proving users' understanding between example-based explanations and visualisation-based explanations — H3a is rejected. In domain 1, the difference in the task prediction between the two treatment groups is larger than that in domain 2. Specifically, effect size in domain 1 is $r = 0.23$ ($p = 0.13$) and in domain 2 is $r = 0.03$ ($p = 0.86$).

There are some discrepancies between domain 1 and 2 when comparing example-based and visualisation-based explanations in terms of trust and satisfaction. In the first domain, **H3b** ($p < 0.001, r = 0.26$) and **H3c** ($p < 0.001, r = 0.28$) are supported. However, in domain 2, **H3b** ($p = 0.1 > 0.05$) and **H3c** ($p = 0.06 > 0.05$) are both rejected. We envision the discrepancies of H3b and H3c may be because prior knowledge of participants could affect them doing the tasks in two different domains. Future work could test this idea further.

As observing no statistically significant difference between example-based and visualisation-based explanations, we then use qualitative analysis to find the limits of both designs and suggest directions to design effective explanations.

Qualitative Analysis

We perform the thematic analysis (Braun and Clarke 2006) from the text written by participants after each mutiple-choice question to know why they selected an option. The text is a response to "Can you please explain why you selected this option?". We followed Nowell et al. (2017) who gave a step-by-step approach for doing trustworthy thematic analysis. Three authors were involved in the qualitative analysis. The first author identified and documented the themes and the codes. Through multiple discussion meetings, two other authors critically analysed the codes and verified them. Finally, we decided on the final codes as in Table 5.

Every participant did the same 10 questions so we have 30 (participants) $\times$ 10 (questions) is 300 (texts) for a condition. Given that we have two treatment conditions and two datasets, we analysed a total of 1,200 texts and each text is assigned to one code or more than one code depending on the number of explanations in that response. Each code is classified as one of: (1) a correct answer (C); (2) a wrong answer (W); or (3) "not enough information" (D). The final analysis includes 1,112 texts after removing 88 texts due to poor quality. We found the following observations, which suggest future improvements.

Use text labels instead of numbers to present categorical variables.. A categorical variable can be shown in numbers or text labels. In Table 6, the majority of wrong codes in HR domain is *W-Linear assumption* (78% and 95%) because most explanations using categorical variables are written in numbers. There was no *linear assumption* codes in the income dataset since all explanations used text labels.

When the labels of categorical features indicate ordinal data, visualise counterfactuals help to reduce the error "linear assumption", making it easier for people to interpret the highest or lowest values. According to Table 6 (HR dataset), 95% (39) of wrong responses happened due to *linear assumption* in the example-based condition; however, we found only 78% (21) of *linear assumption* in the visualisation-based condition. For instance, in a question where the job level is a categorical variable and is not correlated with the confidence score, a participant in the example-based condition mentioned: *"Those with a higher job level had a higher confidence rating"*. In contrast, another participant in the visualisation-based condition could identify the highest confidence value at job level 2 without mentioning about the linear trend: *"The AI predicted job level 2 has the highest chance of staying"*.

It is hard for people to interpret the example-based explanations when the differences between counterfactual outputs and categorical attributes are minimal. According to Table 6 (wrong answer, income dataset, categorical variables), we observe 28% (5) of *W-Small difference*

			Income		HR	
			E	V	E	V
Wrong Answer	Categorical Variables	W-Linear assumption	0 (0%)	0 (0%)	39 (95%)	21 (78%)
		W-Small difference	5 (28%)	0 (0%)	2 (5%)	2 (7%)
		W-Reversed category	13 (72%)	11 (100%)	0 (0%)	4 (15%)
	Continuous Variables	W-Case-based	1 (4%)	0 (0%)	7 (64%)	0 (0%)
		W-Small difference	0 (0%)	2 (18%)	0 (0%)	2 (12%)
		W-Reversed correlation	23 (96%)	9 (82%)	4 (36%)	14 (88%)
Not Enough Information	Categorical Variables	D-Different attribute values	6 (100%)	0	8 (80%)	0
		D-No correlation	0 (0%)	0	2 (20%)	0
	Continuous Variables	D-Outside range	1 (6%)	9 (64%)	2 (13%)	6 (46%)
		D-Different attribute values	4 (24%)	0 (0%)	3 (19%)	0 (0%)
		D-No correlation	12 (70%)	5 (36%)	11 (68%)	7 (54%)
Correct Answer	Categorical Variables	C-Correlation-based	17 (11%)	20 (11%)	18 (10%)	0 (0%)
		C-Case-based	133 (89%)	159 (89%)	157 (90%)	186 (100%)
	Continuous Variables	C-Correlation-based	97 (98%)	118 (100%)	81 (98%)	99 (100%)
		C-Case-based	2 (2%)	0 (0%)	2 (2%)	0 (0%)

Table 6: Frequencies and Percentages of Codes for Explanations

codes in the example-based condition. For example, in a question where *Manager* occupation has the highest confidence score, some participants mistakenly selected *Skilled Specialty* as the highest even though this occupation is the second highest. In this case, the difference in confidence values between *Manager* and *Skilled Specialty* is only 2% (93% and 91%). This small difference made 5 participants chose a wrong answer in the example-based condition.

Using visualisation-based explanations is easier to understand correlations; however, many participants were not willing to extrapolate the correlation beyond the lowest and highest values. In Table 6 (Not Enough Information), we have fewer codes of *D-No correlation* in visualisation-based explanations. However, we record a higher number of codes of *D-Outside range* in this visualisation condition. This issue suggests that we should not expect participants to extrapolate the correlation, and all counterfactual points should be shown in the explanations.

Regardless of variables, if the counterfactual examples in the example-based explanations are not the same as the values in the question, many participants argued that they do not have enough information to decide (*D-Different attribute values*). For example, a participant said: *"Because the position is different, lab tech versus sale rep, I feel that even though the AI chose the one with the highest confidence as the one with the lowest daily rate, I am not sure if the job description would change that confidence level"*. In this question, we provided the example-based explanation for *Sales Representative* job, but the question shows instances of *Lab Tech* job. Even though the daily rate increases linearly in all cases, some participants did not feel confident to apply this observation when we change the instance values in the question. They applied *case-based reasoning* when interpreting the example-based explanation of a linear model rather than interpreting the linear correlation in this explanation. That is, they found the closest example in the counterfactual explanation presented, and compared that example with the question. Similarly, we found an overall 8 codes of **W-Case-based** where participants applied case-based reasoning to do the task with example-based explanations of continuous values. A participant wrote: *"It really is a tough call but I chose employee 1 because the 400 range has the highest percentage of leaving"*. In this example, the participant saw that the daily rate of 400 has the highest confidence of leaving, therefore, they selected the value that is close to 400 rather than interpreting the linear correlation between the daily rate and the confidence score (lower daily rate indicates higher confidence of leaving). Specifically, the question has three daily rate options of 200, 201 and 247, they eventually selected 247 as the final answer, arguing that 247 is closest to 400 in the example-based explanation. In general, **it is clear that participants in the example-based condition used a 'case-based reasoning' approach to understanding the model**. This led participants to overlook the linear trend between the confidence score and the feature values. This finding suggests that we should be careful when using example-based explanations to interpret continuous variables for models, except for cases when the underlying model is itself a case-based model. Using graphs to visualise continuous variables can mitigate this issue.

Conclusion

This paper proposes two approaches for counterfactual explanation of model confidence: (1) example-based counterfactuals; and (2) visualisation-based counterfactuals. Through a human-subject study, we show that the counterfactual explanation of model confidence helped users improve their understanding and trust in the AI model. Furthermore, the qualitative analysis suggests directions of designing better counterfactual explanations. In the future, we plan to perform more extensive user studies to evaluate whether we can improve decision making using such explainability techniques.

Acknowledgments

This research was supported by the University of Melbourne Research Scholarship (MRS) and by Australian Research Council (ARC) Discovery Grant DP190103414: Explanation in Artificial Intelligence: A Human-Centred Approach.

References

Antorán, J.; Bhatt, U.; Adel, T.; Weller, A.; and Hernández-Lobato, J. M. 2021. Getting a CLUE: A Method for Explaining Uncertainty Estimates. In *9th International Conference on Learning Representations ICLR Virtual Event, Austria*.

Atkeson, C. G.; Moore, A. W.; and Schaal, S. 1997. Locally Weighted Learning. *Artificial Intelligence Review*, 11(1): 11–73.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D. S.; Lasecki, W. S.; and Horvitz, E. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33: 2429–2437.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; Nachman, L.; Chunara, R.; Srikumar, M.; Weller, A.; and Xiang, A. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.*, 3(2): 77–101.

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?

Byrne, R. M. J. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, 6276–6282.

Delaney, E.; Greene, D.; and Keane, M. T. 2021a. Instance-Based Counterfactual Explanations for Time Series Classification. In *Case-Based Reasoning Research and Development*, 32–47.

Delaney, E.; Greene, D.; and Keane, M. T. 2021b. Uncertainty Estimation and Out-of-Distribution Detection for Counterfactual Explanations: Pitfalls and Solutions. *CoRR*, abs/2107.09734.

Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; and Das, P. 2018. Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 590–601.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. Accessed: 2023-03-01.

Förster, M.; Hühn, P.; Klier, M.; and Kluge, K. 2021. Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations. In *54th Hawaii International Conference on System Sciences, HICSS*, 1–10.

Goldstein, A.; Kapelner, A.; Bleich, J.; and Pitkin, E. 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1): 44–65.

Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 2376–2384.

Gurobi Optimization, LLC. 2023. Gurobi Optimizer Reference Manual. https://www.gurobi.com. Accessed: 2023-03-01.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR*, abs/1812.04608.

Jacovi, A.; Swayamdipta, S.; Ravfogel, S.; Elazar, Y.; Choi, Y.; and Goldberg, Y. 2021. Contrastive Explanations for Model Interpretability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1597–1611.

Keane, M. T.; and Smyth, B. 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: 28th International Conference, ICCBR*, 163–178.

Kenny, E. M.; Ford, C.; Quinn, M.; and Keane, M. T. 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.

Lai, V.; and Tan, C. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38.

Lewis, D. D.; and Gale, W. A. 1994. A Sequential Algorithm for Training Text Classifiers. In *SIGIR*, 3–12. London.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Miller, T. 2021. Contrastive explanation: a structural-model approach. *Knowl. Eng. Rev.*, 36.

Miller, T.; Howe, P.; and Sonenberg, L. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *CoRR*, abs/1712.00547.

Molnar, C. 2019. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/. Accessed: 2023-03-01.

Monarch, R. M. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 607–617.

Nowell, L. S.; Norris, J. M.; White, D. E.; and Moules, N. J. 2017. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, 16(1).

Pavansubhash. 2017. IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset. Accessed: 2023-03-01.

Riveiro, M.; and Thill, S. 2021. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence*, 298.

Russell, C. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 20–28.

Tomsett, R.; Preece, A.; Braines, D.; Cerutti, F.; Chakraborty, S.; Srivastava, M.; Pearson, G.; and Kaplan, L. 2020. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, 1(4).

van der Waa, J.; Nieuwburg, E.; Cremers, A.; and Neerincx, M. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291.

van der Waa, J.; Schoonderwoerd, T.; van Diggelen, J.; and Neerincx, M. 2020. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, 144.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31.

Wang, D.; Zhang, W.; and Lim, B. Y. 2021. Show or suppress? Managing input uncertainty in machine learning model explanations. *Artificial Intelligence*, 294.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. E. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of Conference on Fairness, Accountability, and Transparency*, 295–305.