Counterfactual Explanations for Misclassified Images: How Human and Machine Explanations Differ

Eoin Delaney [1 2 3] **Arjun Pakrashi** [1 2 3] **Derek Greene** [1 2 3] **Mark T. Keane** [1 2 3]

Abstract

Counterfactual explanations have emerged as a popular solution for the eXplainable AI (XAI) problem of elucidating the predictions of black-box deep-learning systems because people easily understand them, they apply across different problem domains and seem to be legally compliant. While 100+ counterfactual methods exist in the literature, few of these methods have actually been tested on users (∼7%). Even fewer studies adopt a user-centered perspective; for instance, asking people for *their* counterfactual explanations to determine *their* perspective on a "good explanation". This gap in the literature is addressed here using a novel methodology that (i) gathers human-generated counterfactual explanations for misclassified images, in two user studies and, then, (ii) compares these human-generated explanations to computationally-generated explanations for the same misclassifications. Results indicate that humans do not "minimally edit" images when generating counterfactual explanations. Instead, they make larger, "meaningful" edits that better approximate prototypes in the counterfactual class. An analysis based on "explanation goals" is proposed to account for this divergence between human and machine explanations. The implications of these proposals for future work are discussed.

1. Introduction

As Artificial Intelligence (AI) is increasingly used in everyday life for high-stakes decision-making, many new roles have emerged for eXplainable AI (XAI) (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Goodman & Flaxman, 2017; Sokol & Flach, 2019). For instance, in computer vision systems, explanations can help to debug black-box models (e.g., showing why images were misclassified) (Ross & Doshi-Velez, 2018; Bäuerle et al., 2018), to audit system safety (e.g., why a self-driving car misidentified a postbox as a red light (Goyal et al., 2019)), to assess fairness and bias (e.g., why one person's face was cropped from an image over another's (Birhane et al., 2022)) and, even, to provide novel domain insights (e.g., identifying mass legions in digital mammography (Barnett et al., 2021)).

In computer vision, many different strategies have been advanced to explain model predictions (Lipton, 2016; Adadi & Berrada, 2018; Guidotti et al., 2018) using, for instance, saliency maps (Zhou et al., 2016; Selvaraju et al., 2017), feature importance (Ribeiro et al., 2016; Lundberg & Lee, 2017), prototypes (Kim et al., 2016; Rudin, 2019), and factual (Sørmo et al., 2005; Keane & Kenny, 2019), counterfactual (Miller, 2019; Byrne, 2019) or semifactual examples (Kenny et al., 2021; Aryal & Keane, 2023).

Counterfactual explanations have received significant attention in the XAI literature, as they provide "what if" explanations that use a contrasting case to show how a prediction would change *if* the input features had been different (Goyal et al., 2019; Guidotti et al., 2019; Miller, 2019; Karimi et al., 2020; Keane et al., 2021). For image classification tasks, the counterfactual used typically involves making minimal changes to the original instance that flip the original decision. More formally, given a black-box classifier b and I as some to-be-explained query image with the predicted class $b(I) = y$, then I' is a candidate counterfactual explanation when $b(I') = y'$, where y and y' are contrasting classes (see e.g., (Goyal et al., 2019)).

The current AI interest in counterfactual methods has been boosted by philosophical proposals about their centrality in causality (Lewis, 2013; Woodward, 2005), together with psychological findings that they are important to people's understanding of causes (Miller, 2019; Byrne, 2019; 2007; Mueller et al., 2019; Lagnado et al., 2013), and legal analyses suggesting they are GDPR compliant (Wachter et al., 2017). Indeed, there are now 120+ counterfactual methods in the XAI literature, that claim to generate the *plausible* counterfactual explanations people need to understand AI

[1]School of Computer Science, University College Dublin, Dublin, Ireland. [2]Insight Centre for Data Analytics, Dublin, Ireland [3]VistaMilk SFI Research Centre, Ireland. Correspondence to: Eoin Delaney <eoin.delaney@insight-centre.org>.

ICML (International Conference on Machine Learning) Workshop on Counterfactuals in Minds and Machines, Honolulu, Hawaii, USA.

systems (Keane et al., 2021; Verma et al., 2021). However, most of these plausibility claims are based on intuition rather than hard psychological evidence (Barocas et al., 2020; Leavitt & Morcos, 2020). As with much of the XAI literature (Keane & Kenny, 2019; Anjomshoae et al., 2019), user testing of proposed methods is still relatively scarce. Recent work has found that only ∼7% of counterfactual methods specifically were evaluated in this way (Keane et al., 2021).

Accordingly, in this paper, we advance a novel methodology to look more closely at how people *actually* use counterfactuals by asking them to explain images misclassified by an AI system. We then compare their explanations to those generated by benchmark counterfactual methods for the same misclassifications. As these human-generated explanations are, by definition, *plausible* they provide one way to assess the claims made for machine-generated counterfactuals. To presage our results, we find that in these tasks human- and machine-generated counterfactuals are markedly different, that people's counterfactual explanations rely more on prototypes from a contrasting class, rather than minimally-edited instances close to decision boundaries. However, as we shall see, this does not mean that current XAI methods are necessarily *wrong*, although it does show that current methods need to consider the different explanation goals adopted by users in different explanatory contexts.

1.1. Contributions & Outline of Paper

This paper aims to make significant progress in advancing a more user-centered perspective on the use of counterfactual explanations in XAI. We make several novel contributions:

- Providing an up-to-date survey of the main user-study findings on counterfactual visual explanations in XAI including a critical analysis that reveals the system-centered nature of this work.
- Advancing a new user-centered methodology for collecting the counterfactual explanations used by people (2000+ counterfactuals), showing how they can be related to matched explanations from computational, counterfactual methods.
- Finding the divergences that occur between human and machine explanations when evaluation metrics for proximity, representativeness, and prototypicality are applied, accounting for these divergences using the notion of "explanation goals".

2. Current User Testing on Image Datasets

While many counterfactual methods have specifically been proposed for image datasets (e.g., (Chang et al., 2018; Dhurandhar et al., 2018; Goyal et al., 2019; Hendricks et al., 2018; Vermeire et al., 2022)), only a handful of papers consider user testing in computer vision domains. Unfortunately, the few papers that do test image-data have significant issues with their experimental designs, statistical analyses and/or the statistical significance of the results (Goyal et al., 2019; Larasati et al., 2020; Singla et al., 2020; Zhao et al., 2021). So, there are really only three core papers that report anything indicative on the topic (Akula et al., 2020; Cai et al., 2019; Goyal et al., 2019).

Goyal et al. (2019) proposed an influential method, Counterfactual Visual Explanation (CVE), that highlights key regions in an image (e.g., the beak colour of a bird) as feature differences behind counterfactual class changes (e.g., classifying a bird image as an auklet or a cormorant). They performed a user study (N=26)[1] with three conditions testing a no-explanation control against two explanation conditions (i.e., a non-counterfactual feature-region explanation and counterfactual-region explanation). They found the counterfactual-region explanation elicited the highest accuracy (77.8%), followed by the feature-region explanation (74.3%), followed by the no-explanation controls (71.1%), differences that were only significant at lower-than-usual confidence levels (i.e., 87% and 51%). So, at best, these results are indicative rather than conclusive.

Cai et al. (2019) used QuickDraw Doodles (one of the datasets we use here) to reveal more conclusive results in a design that elicited better user interaction. They had participants (N=1,150) generate QuickDraw Doodles of common objects (e.g., draw a helicopter or an avocado) and then had a classifier identify the object using a dataset of labelled drawings. The classifications produced were accompanied by *normative explanations* (i.e., similar examples from the same class, such as other doodles of avocados) or *comparative explanations* (i.e., counterfactuals or similar example from other classes, such as doodles of a pear or potato) with participants being asked to rate how well they understood the system and their views on the system's capability. The results showed that explanations only impacted misclassifications by the system (i.e., no effects for correct classifications) and that the example-based, normative explanations improved people's understanding and assessments of system capability. Unfortunately, these effects did not extend to the counterfactual-based comparative explanations. Cai et al. considered this failure to find counterfactual effects as being due to the "surprisingness" of the counterfactual examples.

Finally, Akula et al. (2020) user-tested their CoCoX method, which adopts a "fault-lines" technique to leverage concepts in creating counterfactuals. CoCoX was compared against CEM-PN (Dhurandhar et al., 2018) and CVE (Goyal et al., 2019), alongside seven other non-counterfactuals methods

[1]For appropriate statistical power this design requires an N>100 and confidence levels should be 95% or higher.

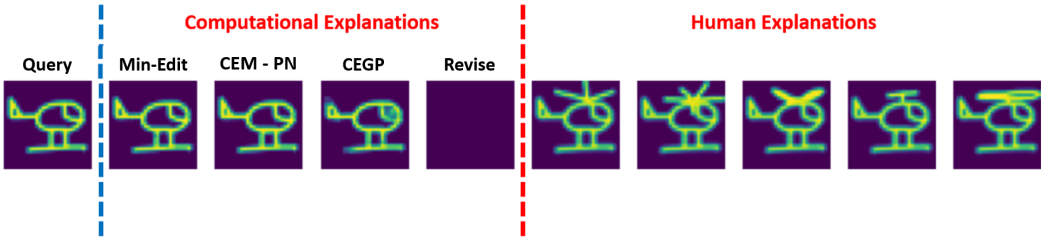


Figure 1. From the QuickDraw data, a query (a "helicopter" misclassified as a "mushroom") and the explanations generated by four XAI methods (as four counterfactual "helicopters") compared to those generated by users. Note, how people add "rotor blades", a semantic-feature, whereas the automated methods perform minimal pixel changes. Revise fails to generate an explanation, confirming findings by Höltgen et al. (2021).

(e.g., LIME, GradCAM, CAM) using two measures: (i) a measure of people's agreement with the model's predictions for test-instances, and (ii) some of the satisfaction questions proposed by the DARPA group (Hoffman et al., 2018). The results showed that CoCoX does best on both measures with CEM-PN and CVE competing for second positions. Furthermore, these explanation conditions do markedly better than no-explanation controls (i.e., 30%-40% better). Although these authors are to be commended for their user-testing efforts, unfortunately this study has several serious design flaws. It appears to be designed as two separate 10-group, between-subjects experiments, one for ML experts (N=20) and one for non-experts (N=60), neither of which are appropriately powered (a 10-condition experiment of this type would require several hundred participants). So, for instance, in the expert experiment this design means that the positive results found for CoCoX are based on just two participants seeing 5 test-items (i.e., 10 data-points), which could by-chance just happen to provide positive results. In addition, five test-materials seem too few. Again, these results on image-datasets are indicative rather than conclusive.

The user studies reviewed thus far have been overwhelmingly *system-centered* ones, in which users are cast as passive recipients of machine-generated explanations. In these studies, XAI methods are used to generate explanations for AI-model outputs, that are then fed to people to be evaluated in different tasks (e.g., for correctness, acceptability, helpfulness, trustworthiness). Such studies lack a reality-check on whether these machine-generated explanations are the ones that people really require. In contrast, a more *user-centered* approach would focus on users, their explanation goals, and their conceptions of the counterfactual explanations in the scenario. Arguably, the user-testing of counterfactual XAI requires a Copernican reversal from being overly system-centered to being more user-centered (see Appendix).

3. A User-Centered Two-Step Methodology

Our methodology realises a human-centered approach in two steps: (i) the collection of human-generated explanations, followed by (ii) comparative evaluations of human- and machine-generated explanations. Two datasets are used: the benchmark MNIST images of written Arabic numbers (LeCun, 1998) and QuickDraw Doodle images (Cai et al., 2019). The latter is arguably more complex than the MNIST one; notably, it involves images with parts that people can readily name (e.g., the toppings on a pizza slice). We train CNNs for each of these datasets and randomly select a sample of misclassifications made by the models. These misclassified instances are then presented to (i) human participants in a psychological experiment and (ii) to each counterfactual method to collect the explanations generated. The two main steps in the methodology are as follows:

- *Human Explanation Collection.* People were provided with a simple editing tool to create their own counterfactual explanations for misclassified images from the CNN for each dataset. This collection was done in two separate experiments, one using the MNIST items (N=42) and a separate pilot study using the QuickDraw items (N=5).
- *Human-Machine Comparative Evaluation*. The same misclassified images were then presented to each of four counterfactual methods – Min-Edit, CEM-PN, CEGP, and Revise – to produce parallel sets of machine-generated explanations, before doing a human-machine comparative evaluation of the sets. We use benchmark evaluation metrics that have previously been used in computational evaluations of counterfactual methods to assess plausibility claims (i.e., on proximity, representativeness, prototypicality).

This methodology tests whether the explanations generated by people correspond to those generated by these counterfactual methods. As such, to the best of our knowledge, this is the first true user-centered assessment of counterfactual

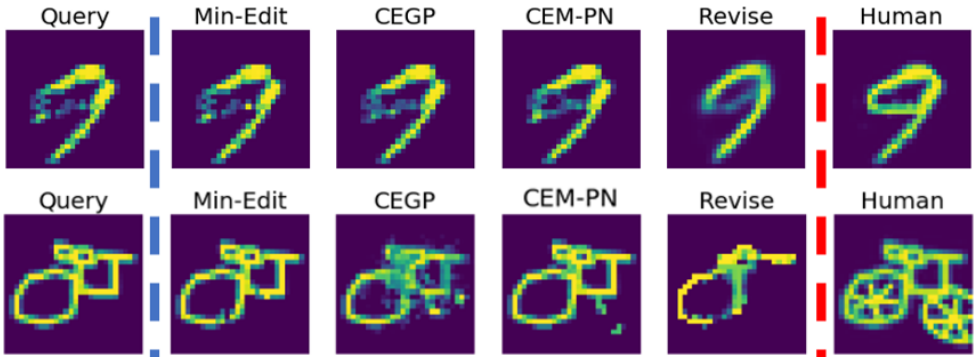


Figure 2. Using the MNIST and QuickDraw datasets, two misclassified query images and their corresponding counterfactual explanations generated by four XAI methods (Min-Edit, , CEM-PN and Revise) and by humans (natural instruction group).

algorithms. In the following sub-sections, we provide more details on each of the steps in this methodology.

Model Setup: CNN Classifier & Datasets: The to-be-explained, black-box model was a convolutional neural network (CNN) trained using a well-known architecture (Van Looveren & Klaise, 2021). Two image datasets were used: MNIST (LeCun, 1998) and Google QuickDraw (Cai et al., 2019). The *MNIST dataset* contains images of written numbers, with 70,000 images covering 10 classes (i.e., the digits 0–9). These images were scaled to $[-0.5, 0.5]$ using the default training and test sets. Dropout layers were implemented for regularization and to facilitate uncertainty computations, using MC-Dropout (Gal & Ghahramani, 2016). The CNN was trained with an Adam optimiser for 10 epochs using a batch size of 256, to achieve an accuracy of 98.93% on the test set, resulting in 107 to-be-explained images being misclassified by the model. The Google QuickDraw dataset contains images gathered from studies that presented people with common objects, asking them to draw the object in 20 seconds as a "doodle". It has 50 million doodle images covering 345 classes (i.e., common object categories such as "bicycle"). The architecture of the classifier was the same as that used on the MNIST dataset. This CNN was trained on a sample of 35,000 images from 5 categories (i.e., "bicycle", "giraffe", "helicopter", "mushroom", and "pizza") using an Adam optimiser for 10 epochs with a batch size of 256. The model achieved an accuracy of 97.02% on the test set, resulting in 447 to-be-explained images which were misclassified.

Materials and Participants: The same misclassifications were presented to people in the user-tests and to the counterfactual methods for the comparative study. For the MNIST dataset, 50 misclassified images were randomly selected from 107 MNIST images misclassified by the CNN. For the QuickDraw dataset, 30 misclassified QuickDraw Doodles were randomly selected from 447 QuickDraw examples misclassified by the CNN. Forty-seven participants took part in the two user studies: the MNIST Study (N=42) and QuickDraw Study (N=5). In the MNIST Study, participants were randomly assigned to two independent groups (see next section for details), the Normal and Min-Edit Groups (both N=21). This sample size was based on a power analysis designed to balance the probability of Type I and Type II errors. Using GPOWER (Erdfelder et al., 1996), for a two separate one-way, t-tests design, with the assumption of a large effect size for each (d = .0.8), the power analysis showed that an N=42 for the overall study ensured an alpha of .05 and power of .80. Both studies were reviewed by an ethics board. Participants were paid an hourly rate of €13.00 in accordance with the living wage in the jurisdiction.

3.1. The Explanation Collection Step: Task & Tool

For the first step in the methodology, a simple software tool was developed to present the misclassified images to participants. The tool allowed images to be edited via a custom interactive GUI implemented using the tkinter Python package. The user data collection was carried out by presenting the CNN's misclassifications to people and asking them to edit the query-image to correct the incorrect prediction (See Appendix 6.1 for full details). For each misclassification, they were told the model's label and its correct label (e.g., 3 and 5, respectively) and that it was misclassified (i.e., "This is an image of a 5 that was incorrectly labelled by the program as being a 3"). They were then invited to edit the image using the editing tool, to explain how the misclassification would have to change to be correctly labelled (i.e., "Your task is to make changes or edits to the image, to help the program correctly label the image as a 5"). This task requires people to create a counterfactual instance that shows how the image would have to change to be correctly classified. Note, we do not have a feedback-loop in this design where the users are provided with live classifier probabilities during editing, as we do not want users to be influenced by potentially-miscalibrated model scores.

In the user test involving MNIST, two separate groups of participants were given slightly different instructions. The "Normal" group was given the instructions discussed above, asking participants to "*...make changes or edits to help the program correctly label the image...*". The "Min-Edit" Group was asked to "*...make the smallest possible changes needed, to help the program correctly label the image...*". This instructional manipulation was designed to determine whether instructions to users to act in accordance with a Min-Edit-type method changed responding relative to the "normal" non-directive instructions.

3.2. The Comparative-Evaluation Step

For the second step in the methodology, the comparative-evaluation step, the counterfactual explanations produced by human and machine were systematically compared using key evaluation metrics that are commonly used in this area. The evaluation metrics used reflect the different perspectives taken on counterfactual goodness and plausibility in the literature, grouping the metrics by (i) *proximity* tests comparing distances between query and counterfactual instances (using L1 and L2 norms), (ii) *representativeness* tests assessing generated counterfactual instances (using MC-Dropout, IM1, 10-LOF, and R% Sub), and (iii) *prototypicality* tests comparing distances generated counterfactual instances to class prototypes. Taken together, these evaluations provide a comprehensive test of divergences/agreements between human- and machine-generated counterfactuals[2].

3.2.1. Counterfactual Methods

Four state-of-the-art counterfactual methods were selected from the literature on counterfactual XAI (Karimi et al., 2021; Keane et al., 2021) based on their (i) popularity as benchmark methods (i.e., according to citations), (ii) their availability as maintained open-sourced code (e.g., on GitHub), and (iii) their ability to handle image data.

Inspired by Wachter et al. (2017); Rips (2010) and implemented using Klaise et al. (2019), the **Min-Edit** method aims to generate a counterfactual explanation by minimizing:

$$(b_t(I') - p_t)^2 + \lambda \|I - I'\|_1 \tag{1}$$

The first loss term pushes the predicted class probability of the candidate counterfactual $b_t(I')$ towards a target p_t, while the second term minimizes the Manhattan distance between the query and counterfactual to promote proximate and sparse solutions, λ, acts as a balancing term.

CEM-PN (Dhurandhar et al., 2018) computes pertinent negatives using an objective function that contains an elastic net ($\beta L_1 + L2$) regulariser to select features to alter via perturbation whilst keeping the perturbations sparse. An autoencoder is leveraged to ensure that the generated explanations lie close to the data manifold through minimizing the L_2 reconstruction error.

Revise (Joshi et al., 2019) relies on a generative model that is a decoder of a variational autoencoder (VAE) trained on the training data. The idea is to minimise the function

$$\ell(b(G(\mathbf{z'})), t) + \lambda \|G(\mathbf{z'}) - \mathbf{I}\|_1 \tag{2}$$

where b is the classifier, t is the target, ℓ is some loss function, and G is the generative model. To find a z' that minimises the loss, z is initialised to the encoding of the original input I. Then the gradient of the loss in the latent space is computed and the algorithm iteratively takes small steps in that space until the prediction changes to the target. Since the resulting counterfactual $I' = G(z')$ is produced by the generative model, it can be *dissimilar* to I in the pixel space. This method is implemented using code from Höltgen et al. (2021), with the recommended hyperparameters of $\lambda = 1$ and gradient step $\delta = 10^{-5}$ and the cross entropy loss for ℓ.

CEGP (Van Looveren & Klaise, 2021) generates a counterfactual by minimising a multi-objective loss function defined by

$$Loss = cL_{pred} + \beta L_1 + L_2 + L_{AE} + L_{proto} \tag{3}$$

where the first term encourages the perturbed instance to belong to the counterfactual class and the elastic net regularizer induces sparse and proximal solutions. The reconstruction error from an auto-encoder is minimised (in L_{AE}) to encourage the counterfactual to belong to the training data distribution. To guide the counterfactual-instance towards the distribution of the perturbed class, the L_2 distance between it and the counterfactual class prototype is minimised in the L_{proto} term. Following the approach of Van Looveren & Klaise (2021), the encoder from L_{AE} is used to retrieve class prototypes.

3.2.2. Evaluation Metrics

Proximity Metrics: The **L1** and **L2** distance metrics are used to evaluate counterfactual methods, measuring the closeness of the counterfactual image, I', to the query image, I, where lower distance-scores are assumed to be a proxy for explanation quality. We compare distance scores for machine-generated query-counterfactual pairs to the corresponding human explanation pairs.

Representativeness Metrics: Monte Carlo Dropout- Following work by Kenny & Keane (2021) and Delaney et al. (2021), we leverage MC-dropout to evaluate counterfactual explanations by estimating the posterior mean of the predictive distribution **MC-Mean** (higher is better) and the posterior standard deviation **MC-Std** (lower is better). The intuition is that explanations with lower uncertainty scores

[2]All code and data available post-review for reproducibility.

should be more representative of the counterfactual class as they are better grounded in the data distribution (Davis et al., 2020).

R%-Substitutability: Inspired by Samangouei et al. (2018) and Kenny & Keane (2021), the generated counterfactuals are used as training data to fit to a 1-NN classifier (in pixel space) which then predicts the full test-set. For MNIST, as we are using 50 instances we compare to an MMD Prototype 1-NN classifier (Kim et al., 2016) that achieves 75.57% accuracy on the full MNIST test set, using only 50 prototypical instances and a Euclidean distance function. A method that achieves half this accuracy would achieve an R% - Substitutability score of 50%.

IM1: Originally presented by Van Looveren & Klaise (2021) as an interpretability metric, using the reconstruction error from a convolutional autoencoder; A lower value of IM1 implies that the candidate counterfactual image I' can be better reconstructed by autoencoders that have seen instances of the counterfactual class, relative to an autoencoder that has seen instances in the original class, implying that I' lies closer to the data manifold of c'.

10-LOF: Following Kanamori et al. (2020), the 10-LOF algorithm (Breunig et al., 2000), is be used to determine if a counterfactual explanation is within the data distribution by computing the local density deviation with respect to its neighbours in the pixel space. The decision-score metric is centred on zero, with higher values indicating that a sample is more within the distribution according to 10-LOF.

Prototypicality Metrics: The **MMD-critic** (Kim et al., 2016) method is implemented to compute prototypes by minimizing the maximum mean discrepancy between the prototype distribution and the data distribution using a kernel density function. This evaluation aims to determine whether generated counterfactuals are actually close to prototypes for the counterfactual class. The **Grad-Cos** metric allows us to determine whether machine-generated and human-generated counterfactuals are close to prototypes in the latent space and is briefly described below. Given some labelled input image $I_A = (x, y)$ and a black-box neural network, $b_\theta(I)$ that is parameterized by θ, with loss $\ell(I;\theta)$ and gradient $\nabla_\theta \ell(I;\theta)$; Grad-Cos (Charpiat et al., 2019) is a gradient based similarity metric that quantifies the degree to which the loss will change when a small update to the model is made using some candidate training instance, I_B, (e.g., a class prototype). If these two images are very similar from the neural network's perspective, this change will be large. Formally, the cosine similarity of gradients kernel can be expressed as:

$$k_\theta(I_A, I_B) = \frac{\nabla_\theta \ell(I_A) \cdot \nabla_\theta \ell(I_B)}{|||\nabla_\theta \ell(I_A)|||||\nabla_\theta \ell(I_B)|||} \quad (4)$$

4. Results: Comparative Experiments

Figure 7 shows some representative data on the types of explanations generated by people and the four methods examined; even a cursory glance at these items shows that the human explanations tend to be more complete and identifiable instances of the counterfactual class for the misclassified instances. Next we explore our results in terms of proximity, representativeness and prototypicality.

Proximity Evaluation: The distance measures for machine-generated query-explanation pairs diverge significantly from the human-generated pairs; human counterfactual explanations are *not* Min-Edits of queries, instead humans make large edits to the query when generating counterfactuals (see Figure 3).

For the MNIST data, a statistical analysis, using a one-way ANOVA, of the distance metrics found a reliable main effect of Group for L1, $F(5,45)= 294.18$, $p < 0.001$, and L2, $F(5, 45) = 291.82$, $p < 0.001$ (see Figure 3a and 3b). Pairwise comparisons between the groups shows that the L1 and L2 scores for three methods (Min-Edit, CEM-PN, CEGP) were all significantly lower than those for humans (all $p < 0.001$; using t-tests and a Bonferroni-Holm correction). In contrast, the Revise method is much closer to the human explanation; on L1 its distance scores are higher than human ones ($p < 0.001$) but on L2 it is not reliably different from the Normal group ($p > .05$). Notably, Even when we explicitly instruct people to act in a Min-Edit way, they do not Min-Edit the images to the same degree as the methods do. *For the QuickDraw data*, the L1 and L2 distance in the pixel space, show essentially the same patterns between groups; a one-way ANOVA found a reliable main effect of Group for L1, $F(3,26) = 107.03$, $p < 0.001$, and L2, $F(3, 26) = 123.62$, $p < 0.001$ (see Figure 3c and 3d).

Representativeness Evaluation: Counterfactual explanations should be within distribution and, to some degree, representative of the counterfactual class. But, how do the within-distribution properties of human explanations compare to those of machine explanations?

The Monte Carlo Dropout (MC-Mean, MC-Std) (Gal & Ghahramani, 2016) metric which measures the uncertainty in a model's prediction confidence (Kenny & Keane, 2021; Delaney et al., 2021; Bhatt et al., 2021), shows that human counterfactuals are the least uncertain with respect to the model's classification (Kenny & Keane, 2021; Gal & Ghahramani, 2016), whereas all four XAI methods have lower certainty scores. Notably, Revise, which was closest to the human counterfactuals on distance, diverges more than any other method on this measure, indicating that its explanations are distributionally quite different to the human ones. In short, humans do not create visual explanations that

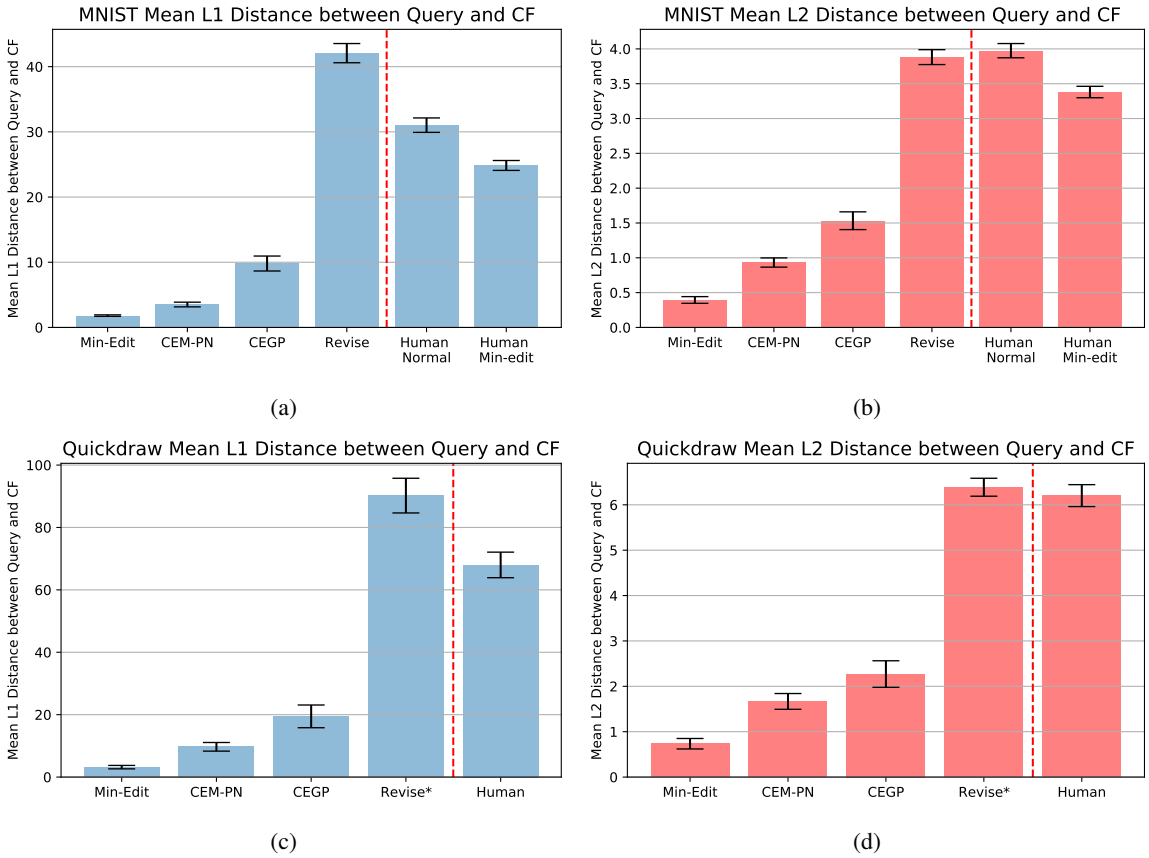


Figure 3. Proximity Evaluations. Mean L1 and L2 distance scores for query-explanation pairs produced by four counterfactual XAI methods – Min-Edit, CEM-PN, , Revise – (left of the dotted red line) compared to the human counterfactuals (right of dotted red line), for (a, b) MNIST and (c, d) QuickDraw datasets (normal instruction group). Error bars show standard error of the mean.*Note, Revise often failed to generate counterfactuals for the QuickDraw data (a coverage deficit also found by Höltgen et al. (2021)); so, Revise's results only reflect instances where explanations were found, making its quite poor performance look better.

Table 1. Representativeness Evaluations. Five out-of-distribution measures for the XAI methods (Min-Edit, CEGP, CEM-PN and Revise) compared to human responses for A - MNIST and B - QuickDraw (bold indicates best score in each case).

	MC-Mean		MC-Std		IM1		10-LOF		R%-Sub	
CF-Method	A	B	A	B	A	B	A	B	A	B
Min-Edit	0.62	0.34	0.33	0.21	1.01	1.06	0.04	0.00	42.72	41.29
CEM-PN	0.59	0.19	0.33	0.13	1.00	1.10	0.04	0.00	43.17	41.46
CEGP	0.66	0.31	0.30	0.21	1.01	1.03	0.08	0.06	49.25	45.85
Revise	0.33	0.16	0.23	**0.03**	1.04	**0.99**	**0.32**	**0.12**	45.76	49.42
Human	**0.94**	**0.71**	**0.11**	0.15	**0.98**	1.02	0.06	0.05	**50.05**	**55.98**

are close to the model's decision boundary (i.e., ones with high aleatoric uncertainty (Schut et al., 2021)). Furthermore, the R%-sub metric (Samangouei et al., 2018) shows that human counterfactuals are more prototypical with respect to the counterfactual class; they have the highest R%-sub scores showing that they are the most representative of the counterfactual class. Finally, the IM1 and LOF metrics confirm this interpretation. IM1 shows that human counterfactuals lie closest to the data manifold of the counterfactual class when compared to the four XAI methods for MNIST. 10-LOF, which is a proximity based out-of-distribution measure in the pixel space, demonstrates that human explanations are more well grounded in the counterfactual class relative to min-edit counterfactuals.

Prototypicality Evaluation: Human explanations reveal a tendency to produce counterfactuals that can be distant from the query, while being close to the prototype(s) of the counterfactual class. For instance, people's counterfactual explanations for misclassified QuickDraw Doodles show semantic-features being added, informed by prototypes in the counterfactual class (i.e., latent features in many CNNs (Kim et al., 2018; Ghorbani et al., 2019; Chen et al., 2020; Zhang et al., 2021)). Figure 1 shows an image of a "helicopter" that was misclassified as a "mushroom", to which people add "rotor blades" to identify it as a "helicopter". In contrast, the XAI methods make small changes to a few pixels that imperceptibly modify the image. In the latent space, human counterfactuals are more similar than all four XAI methods, to the prototypes of the counterfactual class (even when the method purports to use prototypes – See Appendix for details). These results confirm the intuition that people modify the semantic-features of images in producing counterfactual explanations, shaping these explanations relative to the prototypes of the counterfactual class.

5. Concluding Remarks

Recently, many researchers have argued for a more user-centered explainable AI requiring better tests from a user perspective (Miller, 2019; Barocas et al., 2020; Miller, 2023; Lim et al., 2019). In response, we have advanced a new user-centered paradigm where users generate explanations in canonical tasks with a view to comparing them to those generated by counterfactual XAI methods. Our main finding is that human- and machine-generated counterfactuals are markedly different. In the tasks considered, people's counterfactual explanations were shown to rely more on prototypes from a contrasting class, rather than on minimally-edited instances near decision boundaries. Lewis (2013) argued that counterfactuals were the closest possible, minimally-different world to the current one. The present work shows that people compute those minimal differences in a semantic space, rather than in a pixel space, and do so with a view to representative instances of the counterfactual world, rather than the current one, echoing previous findings in cognitive psychology (Lucas & Kemp, 2015; Quillien & Lucas, 2023). However, we believe that an analysis of these results with respect to "explanation-goals" yields a better interpretation of their significance.

Resolving the Divergence between Human & Machine Counterfactuals: The present studies present us with a puzzling divergence between the counterfactual explanations people propose and those computed by counterfactual XAI methods. We believe this divergence can be accounted for thorough analysis of "explanation goals". Conversational theories (Achinstein, 1983; Bromberger, 1965; 1966; Van Fraassen et al., 1980) cast the explanation process as a communicative act between agents with specific explanation goals. These goals shape how an explanation is generated, evaluated, and interpreted by those agents (Sørmo et al., 2005). Most, if not all, current counterfactual XAI methods implicitly assume a task-situation involving a "class-discrimination" explanation goal, in which the counterfactual is designed to communicate discriminating differences between instances; hence, the methods compute minimal (edit) changes to explain things. However, when we pose the same task-situation to people, they seem to implicitly assume a "class-distribution" goal, in which the counterfactual is designed to communicate broad knowledge about classes in the domain; hence, people leverage their knowledge of prototypes to explain things. As such, the present results do not show that that people are *right* and current counterfactual methods are *wrong*. Rather, they show us that XAI-methods and people diverge in their (implicit) choice of explanation goals adopted in the task context. Both choices are appropriate in some situations. There are scenarios in which discriminative-explanations are appropriate (e.g., in the classic recourse scenarios). However, there are also situations where distributional-explanations are appropriate (e.g., in learning about domains).

Limitations & Future Directions: The current studies were conducted using grey scale images of handwritten numbers (MNIST) and hand-drawn everyday objects (QuickDraw), rather than on other commonly-used RGB-image datasets (e.g., CIFAR and ImageNet). One promising line of work for this would be to develop a new editing tool by using text annotations from users in combination with recently developed generative models (Ramesh et al., 2022) to create realistic counterfactual edits. The MNIST study reported here was carefully designed using an appropriate power analysis to test people's generation of explanations. The QuickDraw pilot study could also be extended to mirror this. We envision that future algorithmic developments in counterfactual XAI should account for explanation goals, given the diversity of different task applications in AI.

Closing Comments: Our work promotes a user-centered approach to counterfactual XAI, evaluating the differences between explanations generated by people and machines using popular benchmark comparison metrics from the counterfactual literature. Although the results reveal a marked divergence between the explanations produced by humans and machines, this divergence can be resolved by an analysis of the "explanation goals" used in either context. Computational techniques adopt a "class-discrimination goal", making small edits to the query, whereas humans adopt a "class-distribution goal", making large, semantically-meaningful edits to the query guided by prototypes in the counterfactual class. As such, these findings and the analyses advanced point to new avenues for future research.

Acknowledgements

This publication has emanated from research conducted with the financial support of (i) Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics under Grant Number 12/RC/2289_P2 and (ii) SFI and the Department of Agriculture, Food and Marine on behalf of the Government of Ireland under Grant Number 16/RC/3835 (VistaMilk).

References

Achinstein, P. *The nature of explanation*. Oxford University Press, 1983.

Adadi, A. and Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

Akula, A., Wang, S., and Zhu, S.-C. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2594–2601, 2020.

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Aryal, S. and Keane, M. T. "even if" explanations: Prior work, desiderata benchmarks for semi-factual xai. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI-23)*, 2023.

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3 (12):1061–1070, 2021.

Barocas, S., Selbst, A. D., and Raghavan, M. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 80–89, 2020.

Bäuerle, A., Neumann, H., and Ropinski, T. Training deconfusion: An interactive, network-supported visual analysis system for resolving errors in image classification training data. *arXiv preprint arXiv:1808.03114*, 2018.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Birhane, A., Prabhu, V. U., and Whaley, J. Auditing saliency cropping algorithms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4051–4059, 2022.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: identifying density-based local outliers. In *ACM SIGMOD*, pp. 93–104, 2000.

Bromberger, S. An approach to explanation. *Analytical philosophy*, 2:72–105, 1965.

Bromberger, S. Why-questions. In Colodny, R. (ed.), *Mind & Cosmos*, volume 42, pp. 86–111, Pittsburg, USA, 1966. Pittsburg University Press.

Byrne, R. M. *The rational imagination: How people create alternatives to reality*. MIT press, 2007.

Byrne, R. M. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI-19*, pp. 6276–6282, 2019.

Cai, C. J., Jongejan, J., and Holbrook, J. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 258–262, 2019.

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*, 2018.

Charpiat, G., Girard, N., Felardos, L., and Tarabalka, Y. Input similarity from the neural network perspective. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5342–5351, 2019.

Chen, Z., Bei, Y., and Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Davis, J., Zhu, J., Oldfather, J., MacDonald, S., and Trzaskowski, M. Quantifying uncertainty in deep learning systems - An Amazon Web Services Prospective. In *AWS Prespictive Guidance Report*, 2020. URL https://docs.aws.amazon.com/prescriptive-guidance/latest/ml-quantifying-uncertainty/welcome.html.

Delaney, E., Greene, D., and Keane, M. T. Uncertainty estimation and out-of-distribution detection for counterfactual explanations: Pitfalls and solutions. *arXiv preprint arXiv:2107.09734*, 2021.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pp. 592–603, 2018.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Erdfelder, E., Faul, F., and Buchner, A. Gpower: A general power analysis program. *Behavior research methods, instruments, & computers*, 28(1):1–11, 1996.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. Counterfactual visual explanations. In *ICML*, pp. 2376–2384. PMLR, 2019.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.

Hendricks, L. A., Hu, R., Darrell, T., and Akata, Z. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

Höltgen, B., Schut, L., Brauner, J. M., and Gal, Y. Deduce: Generating counterfactual explanations efficiently. *arXiv preprint arXiv:2111.15639*, 2021.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Kanamori, K., Takagi, T., Kobayashi, K., and Arimura, H. Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *IJCAI-20*, pp. 2855–2862, 2020.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., and Valera, I. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, 33, 2020.

Karimi, A.-H., Schölkopf, B., and Valera, I. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362, 2021.

Keane, M. T. and Kenny, E. M. How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In *Proceedings of the 27th International Conference on Case-Based Reasoning (ICCBR-19)*, pp. 155–171. Springer, 2019.

Keane, M. T., Kenny, E. M., Delaney, E., and Smyth, B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021.

Kenny, E. M. and Keane, M. T. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, pp. 11575–11585, 2021.

Kenny, E. M., Ford, C., Quinn, M., and Keane, M. T. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial Intelligence*, 294: 1–25, 2021.

Kim, B., Khanna, R., and Koyejo, O. O. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pp. 2280–2288, 2016.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Klaise, J., Van Looveren, A., Vacanti, G., and Coca, A. Alibi: Algorithms for monitoring and explaining machine learning models, 2019. URL https://github.com/SeldonIO/alibi.

Lagnado, D. A., Gerstenberg, T., and Zultan, R. Causal responsibility and counterfactuals. *Cognitive science*, 37 (6):1036–1073, 2013.

Larasati, R., De Liddo, A., and Motta, E. The effect of explanation styles on user's trust. In *IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20)*, 2020.

Leavitt, M. L. and Morcos, A. Towards falsifiable interpretability research. In *NeurIPS 2020 Workshop on ML Retrospectives, Surveys & Meta Analyses*, 2020.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lewis, D. *Counterfactuals*. John Wiley & Sons, 2013.

Lim, B. Y., Yang, Q., Abdul, A. M., and Wang, D. Why these explanations? selecting intelligibility types for explanation goals. In *IUI Workshops*, 2019.

Lipton, Z. C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Lucas, C. G. and Kemp, C. An improved probabilistic account of counterfactual reasoning. *Psychological review*, 122(4):700, 2015.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

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

Miller, T. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support. *arXiv preprint arXiv:2302.12389*, 2023.

Molnar, C. *Interpretable machine learning*. Lulu.com, 2020.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*, 2019.

Quillien, T. and Lucas, C. G. Counterfactuals and the logic of causal selection. *Psychological Review*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Rips, L. J. Two causal theories of counterfactual conditionals. *Cognitive science*, 34(2):175–221, 2010.

Ross, A. and Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Samangouei, P., Saeedi, A., Nakagawa, L., and Silberman, N. Explaingan: Model explanation via decision boundary crossing transformations. In *European Conference on Computer Vision (ECCV)*, pp. 666–681, 2018.

Schut, L., Key, O., Mc Grath, R., Costabello, L., Sacaleanu, B., Gal, Y., et al. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pp. 1756–1764. PMLR, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2020.

Sokol, K. and Flach, P. A. Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety. In *SafeAI@ AAAI*, 2019.

Sørmo, F., Cassens, J., and Aamodt, A. Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, 2005.

Van Fraassen, B. C. et al. *The scientific image*. Oxford University Press, 1980.

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pp. 650–665. Springer, 2021.

Verma, S., Dickerson, J., and Hines, K. Counterfactual explanations for machine learning: Challenges revisited. *arXiv preprint arXiv:2106.07756*, 2021.

Vermeire, T., Brughmans, D., Goethals, S., de Oliveira, R. M. B., and Martens, D. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, pp. 1–21, 2022.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harv.J.Law Tech.*, 31:841, 2017.

Woodward, J. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

Zhang, R., Madumal, P., Miller, T., Ehinger, K. A., and Rubinstein, B. I. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11682–11690, 2021.

Zhao, W., Oyama, S., and Kurihara, M. Generating natural counterfactual visual explanations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 5204–5205, 2021.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

6. Appendix

6.1. User Study – Additional Information

Editing Tool: The software tool was developed that allowed images to be edited via a custom interactive GUI implemented using the tkinter Python package (see Figure 4 for a screenshot). The presented image was up-sampled to a 600×600 canvas where it could be edited and the final image was down-sampled to the original 28×28 size. Participants had the option to add pixels, remove pixels or reset the image to its original form if they made a mistake. A log of the stroke information carried out by the user and the final edited image for each presented image was recorded and saved for later analysis.

In both studies, after receiving the instructions and practice trials, participants proceeded through all the presented images at their own pace. The presented set of images was randomly shuffled anew for each participant to control for possible order effects. Each experimental session took ∼15-30 minutes (typically, ∼20 min in MNIST Study and ∼15 min in QuickDraw Study), including the final de-briefing on the rationale for the study. The logs of participant's stroke information and final edited image for each item were all recorded and saved after being suitably anonymised.

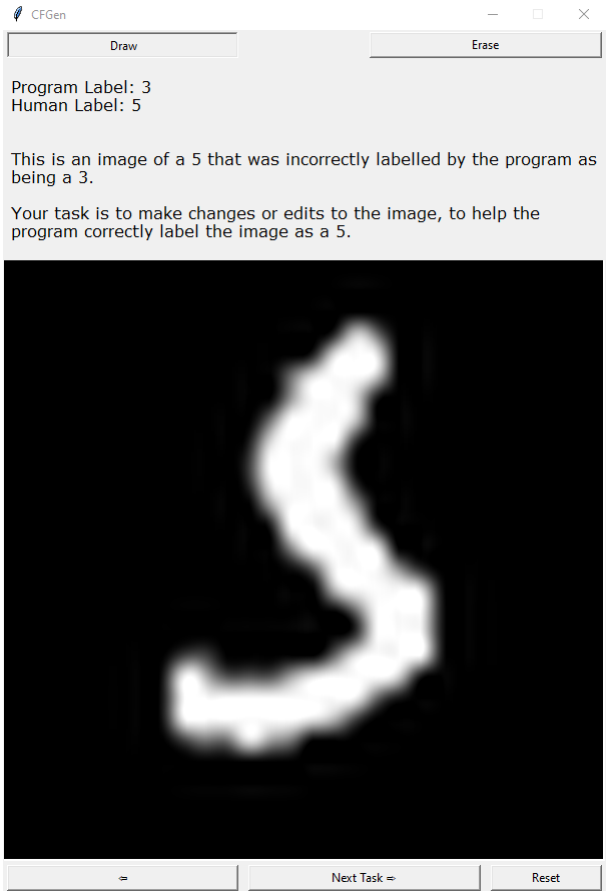


Figure 4. Screenshot of the editing tool used for collecting user explanations, showing a misclassified MNIST image of a "5", along with the instructions to participants. The interface allows pixels to be added or removed using the cursor as a pen or eraser, after clicking the "Draw" or "Erase" buttons, respectively. The "Reset" button removes all edits, resetting the image to its original form.

Response Post-Processing – User Tests: In each of the user studies, for a given misclassified item a response from each participant in the experiment is recorded; so, for MNIST experiment we have 42 explanations for the first misclassification, 42 for the second and so on. So, overall 2,250 human explanations were gathered: 42 people x 50 items for the MNIST experiment and 5 people x 30 items for the QuickDraw experiment. However, for each of the counterfactual methods we have just one explanation per misclassification; so, 320 explanations (80 items x 4 methods). So, to compare the human and machine explanations in a one-to-one fashion, we computed the medoid of human responses to a given item. This group-level response was then used in the explanation-to-explanation comparison for each of the metrics.

6.2. Prototype Evaluations: Prototypes & Similarity

To determine the closeness of generated counterfactuals to the prototype(s) of the counterfactual class, MMD-Critic (Kim et al., 2016) was used to create prototypes for the class and then Grad-Cos was used to measure the latent similarity between explanations and prototypes in order to determine if explanations generated by humans are more similar to class prototypes relative to explanations that are automatically generated. MMD-critic is briefly described below.

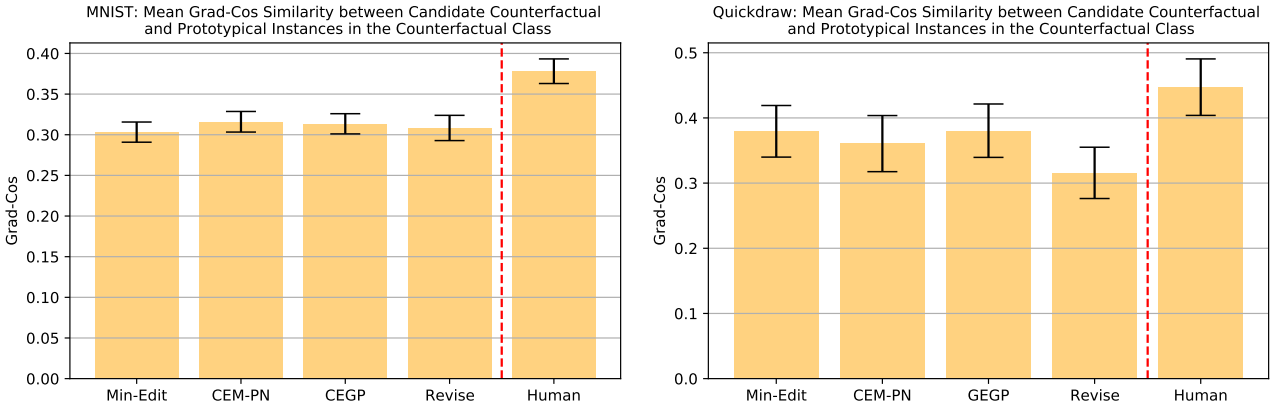
Prototype Retrieval – MMD-Critic: Introduced by Kim et al. (Kim et al., 2016), this approach computes prototypes by minimizing the maximum mean discrepancy between the prototype distribution and the data distribution. These densities are estimated using a kernel density function, k. Following (Molnar, 2020; Kim et al., 2016), let m represent the number of individual prototypes z and n represent the number of data-points x in the dataset. Then the MMD^2 can be represented by:

$$\begin{aligned} MMD^2 = &\frac{1}{m^2}\sum_{i,j=1}^{m} k(z_i, z_j) \\ &-\frac{2}{mn}\sum_{i,j=1}^{m,n} k(z_i, x_j) \\ &+\frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) \end{aligned} \tag{5}$$

The first term calculates the average proximity of the prototypes to each other, while the last term calculates the average proximity of the data-points to each other. The middle term calculates the average proximity between the prototypes and the other data-points (multiplied by 2). In our implementation we use a standard radial basis function as our choice for the kernel k, defined by:

$$k(x, x') = exp\left(-\gamma||x - x'||_2\right) \tag{6}$$

The MMD^2 measure, kernel function and greedy search are combined in an algorithm to find prototypes (Molnar, 2020). Starting with an empty list of prototypes, each point in the class are evaluated using MMD^2, and the point that minimizes MMD^2 to the largest degree is added to the list.



(a) MNIST

(b) QuickDraw

Figure 5. Prototype Evaluations Results. Mean Grad-Cos Similarity scores for counterfactual-prototype and query-prototype pairs (prototypes retrieved using MMD-critic) from XAI methods compared to the human counterfactuals, for (a) MNIST and (b) QuickDraw datasets (Error bars show standard error of the mean).

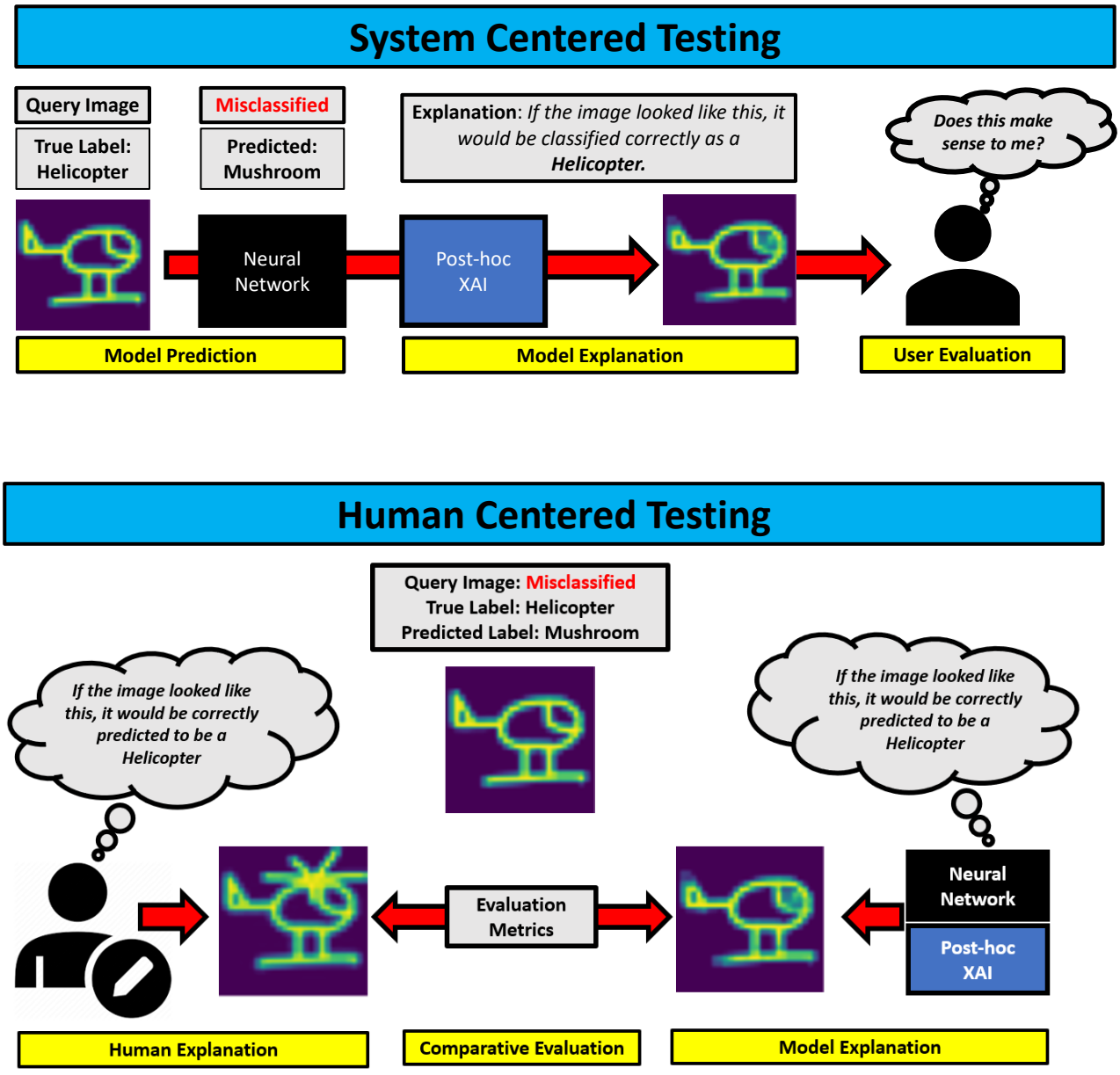


Figure 6. System-centered user-tests of counterfactual XAI present people with the outputs from an AI-plus-XAI method to evaluate them in different ways. User-centered tests try to capture the user's perspective on explanation. In our methodology, an XAI-method's explanation of model outputs (e.g., misclassified images) are evaluated by comparing them to human explanations of the same model outputs (e.g., misclassifications). Most current user-tests of counterfactual XAI are system- rather than user-centred.

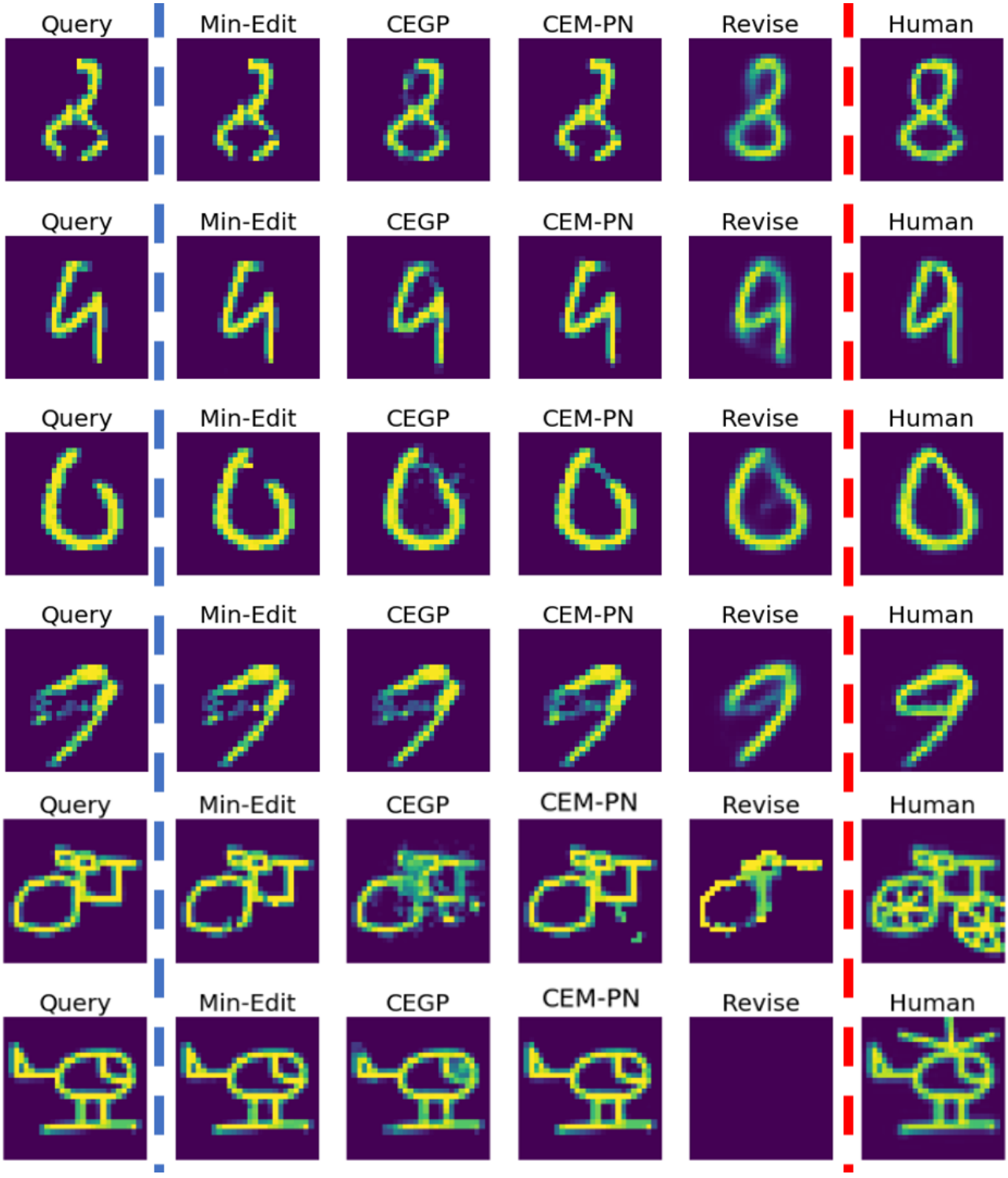


Figure 7. Additional Results Using the MNIST and QuickDraw datasets, misclassified query images and their corresponding counterfactual explanations generated by four XAI methods (Min-Edit, CEGP, CEM-PN and Revise) and by humans (natural instruction group).