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

Eoin Delaney[a,b,c,*], Arjun Pakrashi[a,c], Derek Greene[a,b,c], Mark T. Keane[a,b,c]

[a]*School of Computer Science, University College Dublin, Belfield, Dublin, Ireland*
[b]*Insight Centre for Data Analytics, Belfield, Dublin, Ireland*
[c]*VistaMilk SFI Research Centre, Belfield, Dublin, Ireland*

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. Although over 100 counterfactual methods exist in the XAI literature, each claiming to generate plausible explanations akin to those preferred by people, 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.

Keywords: XAI, Counterfactual Explanation, User Testing

*Corresponding author - **eoin.delaney@insight-centre.org**

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) [2, 35, 43, 105]. For instance, in computer vision systems, explanations can help to debug black-box models (e.g., showing why images were misclassified) [97, 10], to audit system safety (e.g., why a self-driving car misidentified a postbox as a red light [44]), to assess fairness and bias (e.g., why one person's face was cropped from an image over another's [12]) and, even, to provide novel domain insights (e.g., identifying mass legions in digital mammography [8]).

In computer vision systems, many different strategies have been advanced to explain model predictions [83, 2, 48] using, for instance, saliency maps [124, 102], feature importance [96, 85], prototypes [65, 98], and factual [106, 59], counterfactual [89, 18] or semifactual examples [63, 7]. Saliency maps have been extensively used in image classification to highlight "important regions" in the input image by using back-propagation and up-sampling to generate an activation map [124, 102]. In a similar fashion, explanations using model-agnostic feature importance (e.g., LIME [96] and SHAP [85]) aim to show the features or super-pixels of an instance that contributed most to a prediction. Indeed, recently, to achieve better fidelity to the underlying black-box model [3], there has been a move away from input-level features (pixels or segments) to more high-level concepts (e.g., stripes in an image of a zebra) based on extracted latent features (e.g., as concept activation vectors [67, 41, 24, 122]). By the same token, example-based explanations try to leverage meaningful, concept-level units (e.g., such as prototypes, factual, counterfactual or semifactual examples). Here, we focus on counterfactual methods, although our results are also shown to have significant implications for prototype techniques.

Counterfactual explanations have received significant attention in the XAI literature (for reviews see [48, 112, 57, 60]) 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 [44, 47, 89, 57, 60]. For instance, when a two-year loan application for $5k is refused by an automated system and the user asks "Why?", the counterfactual explanation might suggest "If you had asked for a $4.5k loan over a one-year term, then you would have been granted the loan". In this case, changes to the *loan-amount* and *loan-duration* features flip the decision to the user's desired outcome (i.e., *loan-granted*). In image domains, a corresponding counterfactual could be used to audit a misclassification by a black-box deep learner; for instance, when a self-driving car misidentifies a postbox as a red light and the user asks "Why?", the counterfactual explanation might suggest "if the round-red-postbox had been a square-red-postbox, it would *not* have been misidentified as a traffic light". Note, that in both of these cases, the counterfactual used has the same form, involving 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., [44]).

The current AI interest in counterfactual methods has been boosted by philosophical proposals about their centrality in causality [79, 120], psychological findings that they are important to people's understanding of causes [89, 18, 17, 93, 74] and legal analyses suggesting they are GDPR compliant [115]. Indeed, there are now 120+ counterfactual methods in the XAI literature, that claim to generate the *plausible* counterfactual explanations people need to understand AI systems [60, 113]. However, most of these plausibility claims are based on intuition rather than hard psychological evidence [9, 77]. As with much of the XAI literature [59, 5], user testing of proposed methods is still relatively scarce; Keane et al. [60] found that only $\sim$7% of counterfactual methods specifically user-tested the functionalities advanced. So, we really do not know which, if any, of these counterfactual methods *really* generates explanations that people find plausible. Furthermore, arguably, this XAI literature does not pay sufficient attention to the *explanation goals* of users in the tasks examined (see [106] for a detailed analysis of this issue). Specifically, most methods assume that people always and only require minimal-change or Min-Edit counterfactuals to meet their explanatory goals, when this may not always be the case (as we shall see later in Section 3.4).

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*, though it does suggest that current methods should consider using the same explanation goals as those "naturally" adopted by users in different task 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. Accordingly, we make several novel contributions to the field:

- Providing a critical analysis of the notion of plausibility in the counterfactual XAI literature, showing how different intuitive claims for it can be mapped to evaluative metrics measuring proximity, representativeness, and prototypicality.

- Providing an up-to-date survey of the main user-study findings on counterfactual 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, 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", and detailing the major implications they have for current methods and task analyses in the field.

In the next section, we review the related work on counterfactual methods in XAI, their plausibility claims, and the evaluation metrics used to test these claims (see Section 2). Then, we survey the main findings emerging from the scarce user tests that have been performed on the efficacy of counterfactual explanations methods, introducing the proposal that they are predominantly system-centered (see Section 3). We then sketch our novel methodology for collecting human counterfactual explanations for misclassifications in two benchmark image datasets (MNIST [78] and QuickDraw Doodles [19]; see Section 4). Section 5 presents our comparative experiments on counterfactual explanations, detailing the divergences between human-generated and machine-generated counterfactuals. Finally, we discuss the implications of these results, along with their limitations, for counterfactual explanations, in particular and, more generally, for the explanation of AI systems.

2. Related Work I: Plausibility of Counterfactual Explanations

In the current section, we do not attempt to survey the 120+ counterfactual methods that exist in the XAI literature (see [56, 112, 60, 46, 6, 107] for surveys). Rather, we focus on the literature of most relevance to the current study to contextualise the methods tested here. Most counterfactual methods claim, typically on intuitive grounds, to produce *plausible* counterfactuals; that is, counterfactuals that people would find informative, comprehensible and appropriate to their explanation goals. However, researchers differ in how they define plausibility; some identify plausibility with proximity, others with whether the counterfactual is in the data manifold, whereas still others emphasise its prototypicality within counterfactual class. Here, we assess current benchmark methods with respect to all of these different interpretations in an attempt to provide a fair assessment of each.

Consider each of these plausibility perspectives, in turn. Taking a *proximity* perspective, many counterfactual methods since Wachter et al.'s seminal paper [115], argue that plausibility hinges on counterfactual closeness; that counterfactuals close to the query and to the decision boundary of the counterfactual class are the ones people find plausible. This focus on proximity tends to lead to sparse counterfactuals that modify few aspects of the original query instance, so-called "minimal edits", thus making them easier for people to comprehend [61]. From a *representativeness* perspective, other methods propose that plausible counterfactual explanations are in the data manifold and are representative of the domain (e.g., [76]). Finally, from a *prototypicality* perspective, others have argued that plausibility hinges on using semantically-meaningful features that represent the central tendencies of the class, specifically, those captured by *prototypes* (e.g., [110]). In this section, we introduce and critically review the methods tested in the current study in terms of (i) their perspective on plausibility (i.e., proximity, representativeness, prototypicality), (ii) the details of the methods tested, (iii) the evaluation metrics used to assess their particular plausibility perspective.

2.1. Proximity: Plausible Explanations Are Min-Edits

David Lewis' [79] influential philosophical analysis proposed that counterfactuals captured the closest possible world in which some target event minimally differed. In the last few years, this view that plausible counterfactual explanations depend on balancing the distance from the query and decision boundary has, perhaps, been the dominant paradigm in the area. This idea underlies so-called *proximity methods* that aim to find the minimal changes or Min-Edits to a query-instance's features to generate one with a contrasting predictive outcome [61, 89, 92, 110, 115].

The earliest methods in this literature relied on proximity to find Nearest Unlike Neighbours (NUNs) in the dataset as counterfactual explanations [87, 94]. However, these so-called *endogenous* techniques rely on known data-points and existing feature-values and, as such, tend to be overly dependent on the availability of suitably-close instances in the dataset [46]. Wachter al.'s [115] seminal *Min-Edit* method introduced the major innovation of generating *synthetic* counterfactual-instances to provide much better explanatory coverage. However, this *exogenous* technique [46], which uses random perturbation, can lead to other issues (e.g., invalid data-points being proposed as explanations). In passing it should be said that in cognitive psychology plausibility has not always been identified with similarity (see [26],[27] and [60] for a critique).

2.1.1. Testing Proximity: Using Min-Edit

The Min-Edit method searches a space of perturbations of the query-instance using gradient descent, applying a loss function that balances the closeness of the counterfactual to the query against making minimal feature-changes required to deliver a prediction change [115]. So, this 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. The Lagrangian multiplier, λ, acts as a balancing term.

This method has been hugely influential in the recent counterfactual literature and is the recognised baseline for many algorithmic comparisons.

Furthermore, many subsequent counterfactual methods have extended this approach with additional constraints to improve sparsity [29, 45], diversity [92], and to include causal models [57]. Other methods have tried to tackle the invalid-data-point issue by including representativeness ideas, using auto-encoders or generative models to ensure that the counterfactual lies close to the data manifold [33, 54, 64, 103, 110].

In the present comparative study, we use Min-Edit as a baseline for this approach to counterfactual XAI and implement it using [69]. The standard evaluation metrics for assessing plausibility-as-proximity are the L1 and L2 norms, applied to the query-explanation pairings produced by a given explanation method, where lower distances are taken to indicate the success of the method (see Section 4.3.1 for details).

2.2. Representativeness: Plausible Explanations Are Within-Distribution

Part of the motivation behind early counterfactual methods using Nearest Unlike Neighbours (NUNs) was the recognition that explanations would be plausible if they were representative and within the domain. By definition, these methods cannot generate an out-of-domain counterfactual explanation, as NUNs are from the training distribution. However, as we said earlier, if queries have few or no close NUNs, this can still lead to low-quality or failed explanations [46, 61, 104]. Recent perturbation methods (such as Min-Edit) deal with this coverage issue by generating synthetic counterfactuals. However, these methods can risk generating instances that are either unrealistic and out-of-distribution, or not perceptibly different from the query instance [32, 33, 54, 60, 99]. Laugel et al. [76] showed that for some domains these methods could result in 30% of explanations being out-of-distribution. Such concerns led to a re-emphasis on representativeness, to ensure that explanations were plausible by being within-domain and, if possible, not out-of-distribution. This representativeness perspective on plausibility is much closer to proposals made in cognitive psychology, where it is often cast as coherence or consistency with prior knowledge in a domain [25, 26, 27, 39, 70].

2.2.1. Testing Representativeness: CEM-PN & Revise

Many counterfactual techniques have supplemented proximity-based methods to take representativeness into account, often by using auto-encoders or generative models. These methods aim to generate plausible counterfactuals that belong to the data distribution [33, 54, 64]. In the current study, we selected two methods taking this approach – CEM-PN and Revise – based on their recognition in the area, their ability to handle image data, and the open-source availability of their code (see [33, 54]).

CEM-PN [33] 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. This method was implemented using [69].

Revise [54] relies on a generative model that is a decoder of a variational autoencoder (VAE) trained on the training data. The overall 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. Although the L1 norm is known to encourage sparsity, the algorithm may not necessarily generate sparse solutions, as the changes occur in the latent rather than the input space. The method can also fail to generate any explanation at all, in some cases. For instance, [53] reports a 22% failure rate for this method on MNIST images. This method is implemented using code from [53], with the recommended hyperparameters of $\lambda = 1$ and gradient step $\delta = 10^{-5}$. As in [54, 53], we use cross entropy loss for ℓ.

In the present study, these two methods are used as benchmark approaches for this representativeness approach to counterfactual XAI, with all being implemented using [69]. Several different evaluation metrics, typically out-of-distribution measures, have been used to assess plausibility-as-representativeness, applied to the counterfactual-instances produced by a given explanation method. We use MC-Dropout, IM1, 10-LOF and R%Sub to perform these evaluations (see Section 4.3.2).

2.3. Prototypicality: Plausible Explanations are Central to the Class

The final plausibility perspective on counterfactual explanations is the idea that they should be central to the counterfactual class (i.e., that they are close to prototypes). In image domains, because proximity methods try to find counterfactuals which lie just over the decision boundary (minimising distance from the query), they are often not particularly representative of the counterfactual class, or indeed the original query-class. Indeed, as has been discovered in adversarial learning [42], small pixel-changes to a query can sometimes result in a change of class, even when these pixel-changes are actually imperceptible to humans [63, 42] (see Figure 3 for examples). Clearly, such counterfactuals are not plausible explanations for people.

One solution to this problem is to generate the counterfactuals in a latent space using higher-level features (e.g., conceptual [44] or exceptional features [63]). Indeed, some have gone further, arguing that it is better to leverage instances that are maximally representative of a class (i.e., *class prototypes* [65, 23, 66, 80, 8]), or to interpret a model's prediction using human-friendly concepts [67, 41, 24, 122].

2.3.1. Testing Prototypicality: CEGP

In the present tests, we use the popular "Counterfactual Explanations Guided by Prototypes" method (**CEGP**) [110] to implement this approach, even though it was proposed to mainly improve class-representativeness.

CEGP [110] 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. The elastic net regulariser $\beta L_1 + L_2$ aims to ensure sparsity and proximity in the generated instance. L_{AE} is the reconstruction error of the candidate counterfactual instance, which is minimised 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 [110], the encoder from L_{AE} is used to retrieve class prototypes. Specifically, the counterfactual class prototype is defined as the average encoding of the five nearest instances (based on Euclidean distance) in the latent space with the same counterfactual class label. Again, this method was implemented using [69].

To evaluate this plausibility-as-prototypicality perspective we measure the similarity of generated counterfactuals to computed prototypes from the counterfactual class (using Grad-Cos [22] on prototypes from MMD-Critic [65]). MMD-critic [65] is implemented to compute prototypes by minimizing the maximum mean discrepancy between the prototype distribution and the data distribution using a kernel density function (full technical details can be found in Appendix B). This evaluation metric aims to determine whether generated counterfactuals are actually close to prototypes for the counterfactual class. This MMD-Critic & Grad-Cos metric is used to determine whether machine-generated and human-generated counterfactuals are close to prototypes (for details see Section 4.3.3).

3. Related Work II: User Tests of Counterfactual XAI

The recent rapid algorithmic advances in counterfactual XAI have not been paralleled by a significant program of user testing. One survey of user studies reports that only 31% of counterfactual papers (36 out of 127 papers) conducted any form of user evaluation and only 7% competitively test alternative algorithms [60]. Furthermore, many of these studies are vitiated by poor/non-reproducible experimental designs, imbalanced material sets, low/inappropriate numbers of participants, and inappropriate/absent statistical analyses. So, very few studies report solid evidence on the efficacy of counterfactual explanations in XAI. Indeed, as we shall see, most of these user studies adopt a very system-centered approach, in which users are cast as passive recipients of machine-generated explanations. In the following subsections, we critically assess the main findings from the relevant literature under the headings of (i) the general efficacy of counterfactual explanations, (ii) specific comparative tests of methods, (iii) tests of image-based methods, and (iv) the role of explanation goals in counterfactuals.

3.1. Counterfactual Explanations: General Efficacy

Most user studies on counterfactual XAI focus on tabular datasets and, simply, attempt to establish in broad terms whether or not explanations impact user behaviour [34, 81, 84, 117, 46, 88]. Methodologically, they test whether the provision of a counterfactual explanation to an algorithmic decision has some/any effect on user behaviour by comparison to no-explanation controls and/or some other explanatory method (e.g., causal explanations or case-based explanations [59, 116, 117]).

For instance, an early study by Lim et al. [81] gave separate groups of people different explanation-types – What-if, Why-Not, How-to and Why explanations – and found that all of the explanation-interventions improved performance relative to no-explanation controls.However, the Why-Not counterfactual explanations did no better than the other explanation-types. In a similar vein, Dodge et al. [34] assessed four different explanation strategies (e.g., case based, counterfactual, and two forms of global explanation) for biased/unbiased classifiers and found that counterfactual explanations had the greatest impact on user responding (although these tests used quite a small set of materials). Similarly, using a simulated drink-driving-advice app, Warren et al. [116, 117] found that counterfactual explanations improved people's understanding of the app's predictions, over causal explanations and no-explanation controls; they had participants predict outcomes for unseen instances in a testing phase after a training phase in which they were shown the app's decisions with explanations.

However, other studies have found less evidence for the efficacy of counterfactual explanations. In a carefully-controlled study on a simulated app to advise diabetics about insulin treatments, van der Waa et al. [108] found that counterfactual and example-based explanations did not improve people's knowledge of the domain over no-explanation controls, as measured by users' predictive accuracy on test cases. Worryingly, this study also found that people followed the app's advice even when it was incorrect. The latter finding raises the worrying prospect that people can be increasingly satisfied with a system when it is explained to them, without gaining any insight into how the system works (which has been raised as an issue for XAI [16, 116]). Other work [73] that asked people to predict/simulate what a model might do under different inputs found that tasks using counterfactuals elicited longer response times, were rated as being more difficult, and showed reduced accuracy in users. Lucic et al. [84] found that people were less accurate on tasks that asked them to introduce a counterfactual change for an instance than when they were asked to predict outcomes from the input features of the instance. These latter studies point to an added cognitive overhead in processing counterfactuals that has been long-noted in psychology [17, 18].

So, overall, this literature shows mixed support for the efficacy of counterfactual explanations. Of course, part of the problem here may arise from paucity of good studies in the literature. In time, as more studies are done, we may be able to reach a better understanding of what it is that makes counterfactual explanations work in some contexts and not in others. However, it could also be the case that people are not being properly tested in these studies, that the orientation taken is too much from the machine-perspective than a user-perspective.

3.2. Comparative Tests of Counterfactual Methods

Apart from user tests on the general efficacy of counterfactual XAI, given the different proposals on the plausibility of different algorithms, we also require comparative tests between methods. However, even fewer user studies have done such tests [4, 39, 38, 68, 72]. Forster and colleagues [39, 38] have performed a series of unique studies that asked participants to assess counterfactuals produced by different algorithms or parametric-variations of a given algorithm [39, 38]. They compared a gradient-free optimisation method against a popular gradient-based method [115] asking people to rate the typicality, realism and suitability of generated counterfactual explanations (i.e., proxy metrics for representativeness). The results showed that their optimisation method, did better on all metrics than the popular Wachter et al. [115] perturbation method. Kuhl et al. [72] also compared methods that differed in how they selected counterfactual instances (i.e., perturbation methods with and without density computations) but found no evidence that people could distinguish them. Akula et al.'s [4] report a user study that compared their own CoCoX counterfactual method to two other counterfactual methods (i.e., CEM-PN [33] and CVE [44]) showing that their method performed best, though several methodological issues undermine these tests (see next subsection). Finally, Kirfel & Liefgreen [68] specifically tested for effects of "actionable features" on counterfactual explanation (i.e., features that the user is known to be able to change). They found that people's perception of the quality and comprehensibility of explanations was affected by the presence of actionable and mutable features, as opposed to immutable ones. However, they also suggested that the actionable/mutability distinctions made in the methods were not clear-cut in user tests and were often hard to define. So, in summary, the few comparative tests that have been carried out perhaps show less support for benchmark Min-Edit methods in favor of algorithms that are more concerned with computing within-domain explanations, though again the evidence is patchy.

3.3. User Testing on Image Datasets

Only a handful of papers consider user-tests of counterfacual explanations for image datasets, as most focus on tabular datasets. Yet, many counterfactual methods have specifically been proposed for image datasets (e.g.,[21, 33, 44, 50, 114]). 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 [44, 75, 103, 123]. So, there are really only three core papers that report anything indicative on the topic [4, 19, 44].

Goyal et al. [44] reported 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. [19] 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, but the failure could also have been due to other factors (such as the task demands to identify doodles from

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

multiple classes). In passing, it should be said that this study is quite close to the current one, in that it pioneers the user-generation of QuickDraw images in an XAI user-study; however, it differs from the current study in that it asked users to generate training-instances rather than *specifically* asking them to generate counterfactual explanations for test-instances[2].

Finally, Akula et al.'s user-tested their CoCoX method, which adopts a "fault-lines" technique that feature-highlights counterfactual explanations akin to those advanced by CVE [44]. CoCoX was compared against CEM-PN [33] and CVE [44] alongside seven other non-counterfactuals methods[3] (e.g., LIME, GradCAM, CAM) using two measures: (i) a measure of people's agreement with the model's predictions for test-instances (they call this Justified Trust), and (ii) some of the satisfaction questions proposed by the DARPA group [52]. 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 (i.e., 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; assuming all groups received the same five test-items, we do not know whether these particular test-instances just happened to resonate well with training on CoCoX's explanations over other methods (i.e., a larger set of training and test-instances or randomly

[2]To labor this point, Cai et al.'s system retrieved labelled-drawings and used them as counterfactual explanations. Although the original drawings came from people, the AI system selected which drawings to use as counterfactuals, rather than the users. So, people did not select/generate these instances to be used as *explanations*.

[3]The training phase in these experiments appears show people the model's class-selection for 15 test-instances and then asks them for a counterfactual-class for each prediction, with no feedback. This is an unusual training task must hamper transfer-of-training to non-counterfactual methods in the test phase of the study, perhaps explaining why the non-counterfactual methods do so badly.

selected training- and test-instances should have been used). So, again, these results on image-datasets are less than conclusive.

3.4. Counterfactuals & Explanation Goals [4]

The mixed findings emerging from these user-tests of counterfactual XAI methods suggest that something may be missing from current analyses of use-cases in the literature. A candidate for this missing element may be a more thorough consideration of the notion of *explanation goals* (e.g., [82],[106]). Sørmo et al. [106] have pointed out that some philosophers of science emphasise the conversational nature of explanation, where an explanation is seen as a request by an agent to understand something [1, 14, 15, 109]. Taking this perspective, the act of explanation is a communicative process between agents operating within some context guided by explanation goal(s), where the meaningful interpretation of the explanation depends on observing these conditions. So, faced with the scenario, users could have different explanation goals that make one explanation preferable to another. For instance, an accident investigator trying to understand the cause of an automobile crash could prefer a mechanical explanation (e.g., the road surface was slippery) given their goals, while a close friend of the victim could prefer one relying on the driver's mental state (e.g., Jim was very depressed at the time of the accident) given their goals.

Arguably, XAI has not paid sufficient attention to this goal-based perspective on explanation, and counterfactual XAI is no different. Most researchers (implicitly) assume a "class-discrimination" goal, where the counterfactual explanation is aimed at helping users to discriminate the key-differences between instances that change predictions [44, 80, 115] (e.g., the beak-color that discriminates two bird species or the income-difference that changes a loan outcome). Other researchers appear to (implicitly) assume a "class-distribution" goal aimed at helping users to acquire broad knowledge about a domain, examining how features jointly contribute to the predictions made [19, 38, 116, 108]. The class-discrimination explanation-goal holds in classic algorithmic-recourse situations (e.g., the bank loan scenario), where the explanation needs to convey those minimal changes that would yield a preferred outcome. However, the class-distribution explanation-goal occurs in domain-learning situations, where the explanation needs to convey the over-

[4]We thank the reviewers and editor for raising the issue considered in this sub-section.

all "shape" of a domain and its classes (e.g., that different values for weight, gender and units-drunk all contribute in different ways to someone being over the legal blood-alcohol limit for driving).

Most current counterfactual methods seem to be effective in meeting the requirements for class-discrimination explanation goals, though they may not be as "good" at meeting the requirements of class-distribution goals, hence leading to failures in their efficacy in domain learning scenarios (see e.g., [19, 72, 108]). For present purposes, we raise these distinctions as they become important later to the interpretation of our results (see Section 6). In the meantime, it is enough to note that these different explanation goals play a role in situations where counterfactual explanations are used.

3.5. Interim Conclusions on User Testing

In the previous sections, we have seen how the extant user-testing literature on counterfactual explanations has produced mixed and, sometimes, inconclusive results. Indeed, there are further recent results that seem to show that other external factors (i.e., variables independent of counterfactual manipulations) have more impact; factors such as (i) expertise or familiarity with the domain [37, 20], (ii) whether explanation tasks are framed as predictions or diagnoses [28], and (iii) whether instances involve categorical or continuous features [117, 118]. To us, these findings suggest that the community's approach may be overly system-centered rather than being sufficiently human-centred (e.g., insufficient analyses of explanation goals). Also, taking such a perspective may shed a very different light on our current understanding of how people understand counterfactual explanations. In the next section, we consider a new methodology that attempts to rectify these failings.

4. A User-Centered Methodology for Counterfactual XAI

The user studies reviewed thus far have been overwhelmingly *system-centered* ones, in which users are cast as passive recipients of machine-generated explanations[5]. 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

[5]Kaushik et al. [58] is the only user-centered work on counterfactuals that we have found; it taps people's counterfactual understandings of sentiment in texts, but differs from the present work in its focus on text-based counterfactuals for data augmentation.

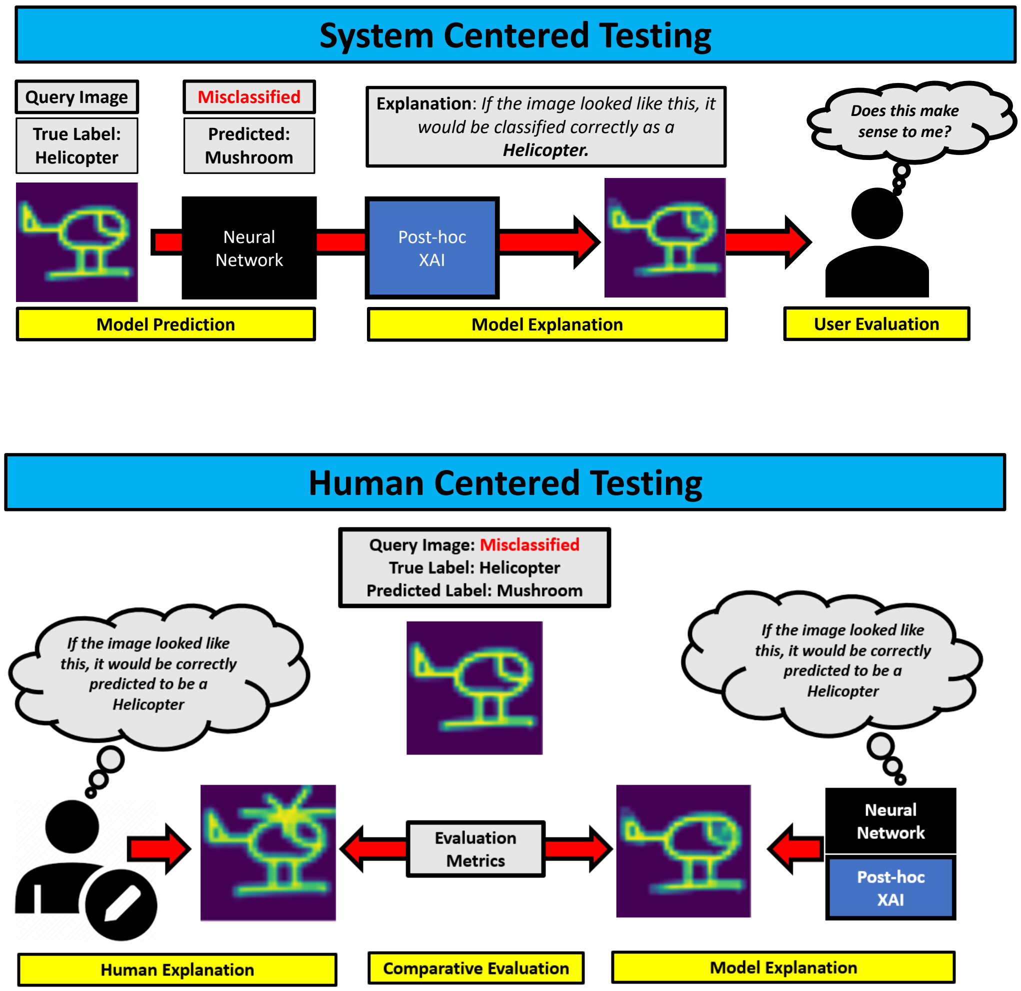


Figure 1: System-centered user-tests of counterfactual XAI present people with the outputs from an AI+XAI method to evaluate them in different ways (e.g., does it help me understand the AI system?). Human-centered tests try to capture the user's perspective on explanation and how they "naturally" approach it in some task-goal context (e.g., capturing their explanation goals and the expectations/knowledge they bring to the task). In our user-centered methodology, an XAI-method's explanation of model outputs (i.e., misclassified images) are evaluated by comparing them to human explanations of the same model outputs (i.e., the same misclassified images). Most current user-tests of counterfactual XAI are system- rather than human-centred. Accordingly, many XAI methods may be psychologically sub-optimal; for instance, if a given method has different explanation goals than users, it is less likely to produce "good" explanations.

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 e.g., Figure 1). However, to date, the XAI community has clearly found it hard to design such user-centered studies. Here, we advance a novel user-centered methodology that starts with gathering user-generated counterfactual explanations for a CNN's misclassifications, before comparing these human-generated explanations to the machine-generated ones from several benchmark counterfactual methods. In the following sections, we introduce (i) this new two-step methodology, (ii) the editing tool used to collect people's explanations, and (iii) the evaluation metrics used to compare the human-generated and machine-generated explanations.

4.1. A User-Centered Two-Step Methodology

The current 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 [78] and QuickDraw Doodle images [19]. The QuickDraw dataset 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 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 explanation 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 algorithms. In the following sub-sections, we provide more details on each of the steps in this methodology.

4.2. 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 (see Figure 2). 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. 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 wanted a "clean" record of what people do in the task without them being influenced by potentially-miscalibrated model scores (though future work could consider such designs).

In the user test involving the MNIST dataset, two separate groups of participants were given slightly different instructions, the so-called "Normal" and "Min-Edit" groups. 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...*". As we saw in Section 2, a key assumption of many algorithms is that

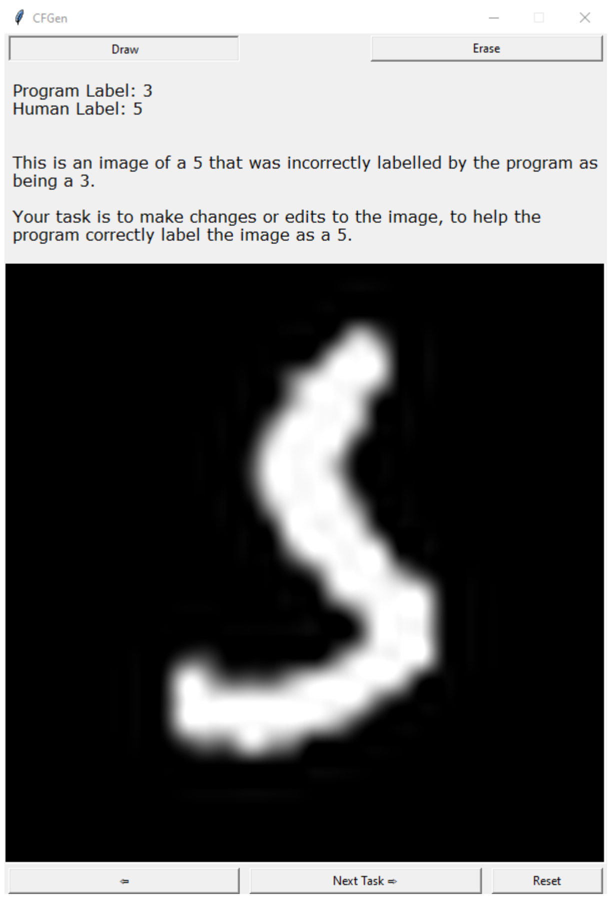


Figure 2: Screenshot of the editing tool used for collecting user explanations, showing a misclassified MNIST image of a "5", along with the instructions to participants. The interface allows pixels to be added or removed using the cursor as a pen or eraser, after clicking the "Draw" or "Erase" buttons, respectively. The "Reset" button removes all edits, resetting the image to its original form.

the counterfactual should make "minimal changes" to the query. 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.

4.3. 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 plausibility in the literature, grouping the metrics by (i) *proximity* tests comparing distances between query and counterfactual instances (using L1 and L2 norms, see Section 4.3.1), (ii) *representativeness* tests assessing generated counterfactual instances (using Monte Carlo Dropout, IM1, 10-LOF, and R% Sub, see Section 4.3.2), and (iii) *prototypicality* tests comparing distances generated counterfactual instances to class prototypes (using Grad-Cos on prototypes retrieved from MMD-Critic, see Section 4.3.3). Taken together, these evaluations provide a comprehensive test of divergences/agreements between human- and machine-generated counterfactuals. In the following subsections, we describe the specific metrics used.

4.3.1. Proximity Metrics: L1 & L2 Norms

The distance metrics – **L1** (Eqn. 4) and **L2** (Eqn. 5) norms – have typically been 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 and plausibility (but see [60] for criticism). Here, we compare distance scores for machine-generated query-counterfactual pairs to the corresponding explanation pairs produced by people in two user studies (see the MNIST and QuickDraw experiments in Section 5.1 and 5.2).

$$\sum_i |I_i - I'_i| \tag{4}$$

$$\sqrt{(\sum_i |(I_i - I'_i)|^2)} \tag{5}$$

4.3.2. Representativeness Metrics: Out-of-Distribution (OOD) Measures

Point representativeness and out-of-distribution metrics – Monte Carlo Dropout [40, 64, 32], R%-Substitutability [100, 64], LOF [13, 55, 31], IM1 [110, 64] – have often been used to evaluate counterfactual methods. These measure the representativeness of explanations in the data-distribution or classes found in the distribution. Here, we briefly describe these metrics that are used to compare the human- and machine-generated counterfactuals.

Monte Carlo Dropout. Originally proposed by [40], MC-Dropout enables one to quickly estimate the uncertainty associated with a prediction by inspecting the variance of a predictive distribution [30]. Multiple stochastic forward passes with different dropout configurations can yield this predictive distribution. Following [64, 32] 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 should be more representative of the counterfactual class as they are better grounded in the data distribution. Full technical details of MC-Dropout and the implementation can be found in Appendix A.

R%-Substitutability. Inspired by [100, 64], the generated counterfactuals are used as training data to fit to a k-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 [65] that achieves 75.57% accuracy on the full MNIST test set (10,000 images), 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 [110] as an interpretability metric, a convolutional autoencoder is trained on the predicted class c, (AE_c) and on the counterfactual class c', $(AE_{c'})$ and the l_2 reconstruction error is monitored to compute IM1. Specifically:

$$IM1 = \frac{\| |I' - AE_{c'}(I')| \|_2^2}{\| |I' - AE_c(I')| \|_2^2} \tag{6}$$

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. Originally presented by [13], the local outlier factor (LOF) algorithm is a distance-based technique that determines whether an instance is out of distribution by computing the local density deviation with respect to its neighbours in the pixel space. Following [55], the 10-LOF can be used to determine if a counterfactual explanation is within the data distribution. The decision-score metric is centred on zero, with higher values indicating that a sample is more within the distribution according to 10-LOF.

4.3.3. Prototypicality Metric: MMD-Critic & Grad-Cos

MMD-critic [65] is implemented to compute prototypes by minimizing the maximum mean discrepancy between the prototype distribution and the data distribution using a kernel density function, full technical details can be found in Appendix B. This evaluation aims to determine whether generated counterfactuals are actually close to prototypes for the counterfactual class. Recall, Min-Edit methods try to find instances that are close to the query but just over the decision boundary in a contrasting class; hence, they are much less likely to be representative members of this counterfactual class. The Grad-Cos metric allows us to determine whether machine-generated and human-generated counterfactuals are close to prototypes and is briefly described below.

Latent Space Similarity: Grad-Cos. 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 [22] 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:

$$k_\theta(I_A, I_B) = \frac{\nabla_\theta \ell(I_A) \cdot \nabla_\theta \ell(I_B)}{|||\nabla_\theta \ell(I_A)||| \, |||\nabla_\theta \ell(I_B)|||} \tag{7}$$

Recent research independently evaluated several popular relevance metrics (e.g., Influence Functions [71]) and found Grad-Cos to work well in practice [49]. Grad-Cos passed the weight randomization tests of [3], indicating that it is faithful to the underlying machine learning model.

5. A Comparative Study: Human & Machine Explanations

As described in Section 4, a two-step, user-centered methodology was applied to the explanation of a CNN's misclassifications after training on two datasets (the MNIST and QuickDraw Doodles datasets). Two separate user experiments, one for each dataset, were carried out to gather human explanations for each of the selected misclassifications. Machine-generated explanations were generated for each of the same misclassifications using four benchmark counterfactual XAI methods – Min-Edit, CEM-PN, CEGP, and Revise (see Section 2). These human-generated and machine-generated explanations where then compared using benchmark evaluation metrics for proximity, representativeness and prototypicality. The following sub-sections lay out the method for this overarching comparative study and specific methods used in the two user-tests that collected the user explanation data.

5.1. ***Method: Comparative Study***

5.1.1. Model Setup: CNN Classifier & Datasets

The to-be-explained, black-box model was a convolutional neural network (CNN) trained using a well-known architecture [110]. Two image datasets were used: the MNIST [78] and Google QuickDraw [19] datasets. 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 $[-0.5, 0.5]$ and the default training and test sets were used. Dropout layers were implemented for regularization and to facilitate uncertainty computations, using MC-Dropout [40]. 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. Pretrained model weights are provided to aid reproducibility. 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" or "helicopter"). 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.

5.1.2. Materials: Comparative Study

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 images misclassified by the CNN.

5.1.3. *Explanation Methods: Comparative Study*

Four state-of-the-art counterfactual methods were selected from the literature on counterfactual XAI [56, 60] 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. The selected benchmark methods, previously outlined in Section 2, are: Min-Edit [115], CEM-PN [33], CEGP [110], and Revise [54].

5.1.4. *Procedure: Comparative Study*

The two user tests were run as independent experiments on each of the datasets, one for MNIST and one for the QuickDraw data. The responses from these user studies were post-processed to find the mean group responses for each misclassified item. The same misclassifications were presented to the four counterfactuals methods and the generated explanations recorded. Pairwise comparisons between the human explanation for a given misclassification and the machine-generated explanation for that misclassification made by each of the four counterfactual methods. These pairings were then evaluated with respect to the six metrics for assessing different aspects of plausibility (see next subsection). Finally, statistical summaries of the results from the metrics were collated and analysed using statistical tests where appropriate.

5.1.5. *Evaluation Measures: Comparative Study*

Eight evaluation metrics, in three groups, were used to process the pairings of human and machine explanations (for detailed descriptions see Section 4.3). The metrics were (i) L1, (ii) L2, (iii) MC-Mean, (iv) MC-Std, (v) IMI, (vi) 10-LOF, (vii) R% Sub and (viii) Grad-Cos Similarity.

5.2. ***Method: User Tests***

5.2.1. *Participants and Design: User Tests*

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, 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 [36], 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. The QuickDraw Study was a single-group pilot to determine whether the findings for the MNIST data generalised to the, arguably, more-complex QuickDraw dataset. Both studies were reviewed by the university's ethics board (ref. LS-E-21-215-Delaney-Keane). Participants were Computer Science students at UCD and were paid an hourly rate of €13.00 in accordance with the living wage in the jurisdiction.

5.2.2. Apparatus: User Tests

The software tool was developed that allowed images to be edited via a custom interactive GUI implemented using the tkinter Python package (see Figure 2 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.

5.2.3. Procedure: User Tests

In both of the user experiments, all participants were tested in a face-to-face experiment with a single experimenter (ED) who presented them with task instructions and the editing tool. After reading the instructions, participants were given three practice trials to learn how to use the tool. The software tool presented one misclassified image at a time, along with (i) a statement about the label it was given by the program and its correct label given by humans, (ii) a statement about how the image was incorrectly labelled by the program, and (iii) the explanation-task instructions to edit the image to help the program correctly label the image. This task asks the user to counterfactually explain the program's misclassification to help improve its subsequent learning. In the MNIST Study, the Normal Group were instructed to "*...make changes or edits to the image, to help the program correctly label the image...*", whereas the Min-Edit Group were instructed to "*...make the <u>smallest possible changes</u> needed, to help the program correctly label the image...*". The latter instructions encouraged this group to conform to a key assumption made by many XAI methods (e.g., the Min-Edit assumption), even though this requirement seems somewhat artificial in this context. In the QuickDraw Study, all participants were given the Normal-Group instructions to "*...make changes or edits to the image, to help the program correctly label the image...*".

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.

5.2.4. 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.

5.3. ***Results & Discussion: Comparative Study***

Figure 3 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 number. These explanations collected from the two user-tests and those generated by each of the four methods were compared using the evaluation measures. The results are grouped and reported for each set of metrics measuring proximity, representativeness and prototypicality. In summary, they show:

- *Proximity Evaluations*: 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 Section 5.4).

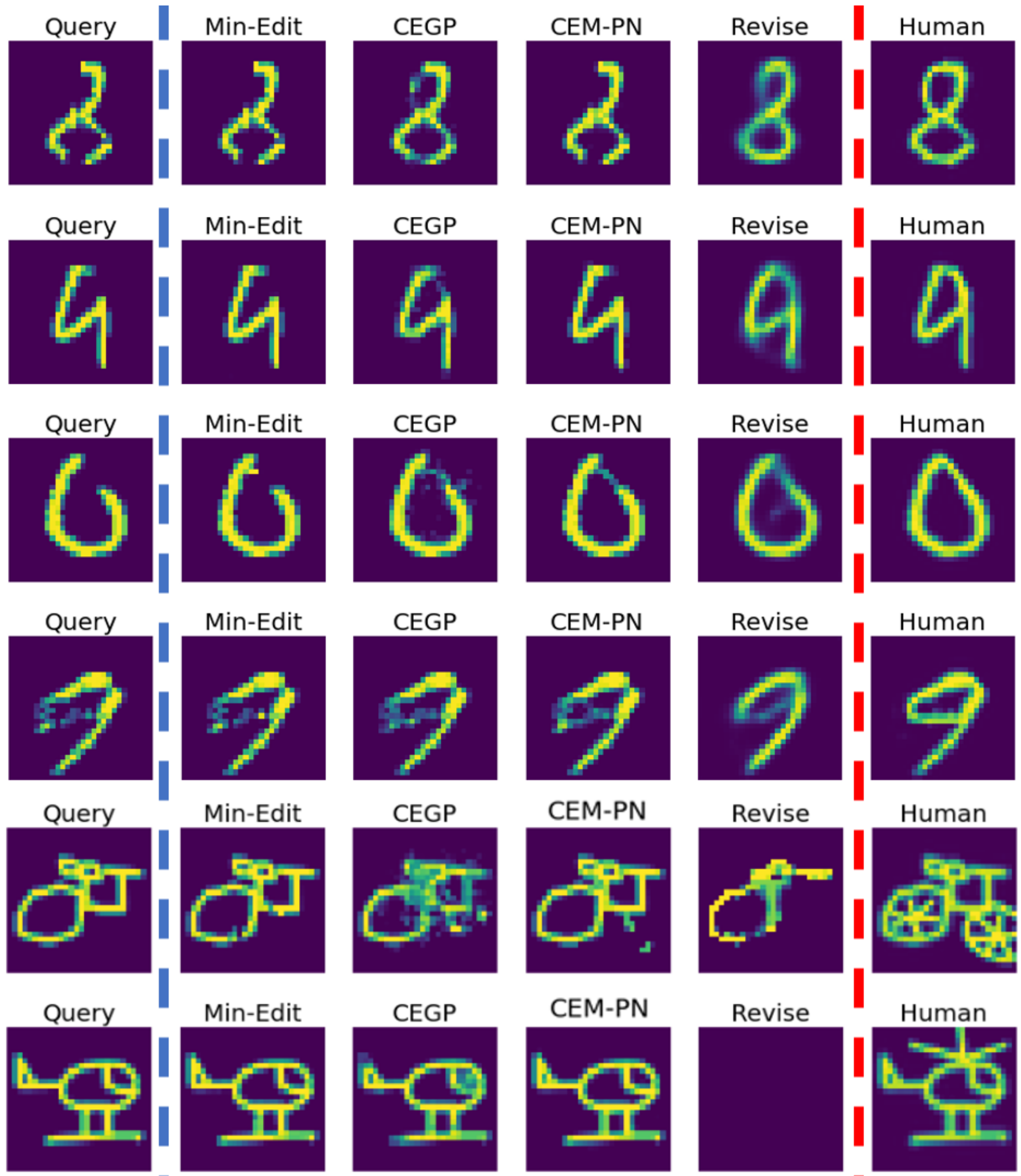


Figure 3: Using the MNIST and QuickDraw datasets, six misclassified query images and their corresponding counterfactual explanations generated by four XAI methods (Min-Edit, CEGP, CEM-PN and Revise) and by humans (natural instruction group).

- *Representativeness Evaluations*: representativeness tests reveal that people's counterfactuals are much more representative of the counterfactual class than the machine-generated explanations; indeed, they are further from the decision boundary and closer to the centre of the counterfactual class distribution (see Section 5.5).

- *Prototype Evaluations*: human counterfactual explanations are much closer to prototypes, computed in the latent space; they tend to sparsely modify high-level semantic features that are not parallelled by the methods, even when the method purports to use prototypes (i.e., CEGP; see Section 5.6).

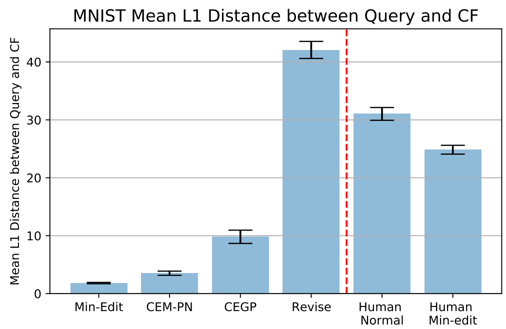
In the following sub-sections we elaborate on the detailed evidence for these summary findings.

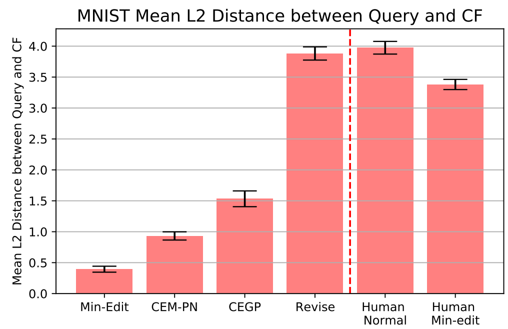
5.4. Proximity: Human Counterfactuals are Not Min-Edits

Popular proximity-based counterfactual methods aim to produce Min-Edit counterfactuals that minimise the distance between the query and explanation instances, achieving a class change between their predictions [115, 119, 92, 110, 29]. So, from this perspective, plausible counterfactual explanations are ones that deliver lower L1 or L2 distances between the query and counterfactual. Figure 4 shows that the distance profiles using L1 and L2 norms for the automated methods, all differ from the human profiles in tests using the MNIST and QuickDraw data. Figure 4a and 4b show the results for the two conditions in the MNIST study (Normal and Min-Edit) which indicate that instructions to people to "minimally edit" impact distance-measures but do not change the overall pattern of divergence between human and machine counterfactuals (see also Appendix C).[6]

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 4a and 4b). 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

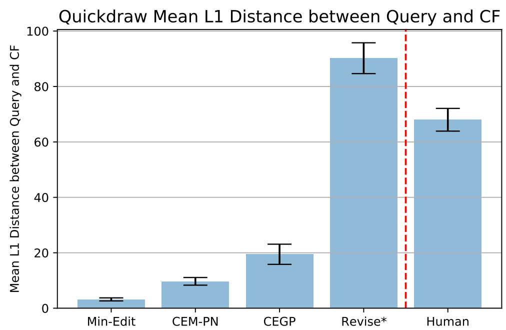
[6]N.b., For the remaining metrics reported, to compare like-with-like, we only use the Normal group's responses, as the Min-Edit group receive non-standard instructions.

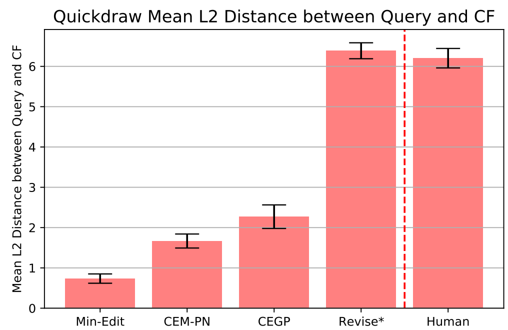




(a)

(b)





(c)

(d)

Figure 4: *Proximity Evaluations*. Mean L1 and L2 distance scores for query-explanation pairs produced by four counterfactual XAI methods – Min-Edit, CEM-PN, CEGP, 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 failed to generate counterfactuals for many instances in the QuickDraw dataset (a coverage deficit also found by [53]); so, Revise's results only reflect instances where explanations were found, making its quite poor performance look better.

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 ($p > .05$).

Figure 3 provides a representative sample of data from the MNIST user-test, revealing how the Revise and human counterfactuals tend to be well-formed examples of the counterfactual class (intuitively, they appear almost prototypical). In contrast, the other three methods produce counterfactuals that are minimal changes to the original query, that are not representative members of the counterfactual class.

Notably, this pattern of results does not change, even when participants were specifically instructed to use a Min-Edit strategy. Recall, in the MNIST user-test, half of the participants were specifically instructed to minimally-edit the images in contrast to more "normal" instructions (Min-Edit versus Normal groups). Although, these instructions significantly reduced the distance scores in the Min-Edit group relative to the Normal group (t-tests, all $p < 0.001$), the human distance scores for query-explanation pairs were all still significantly different and higher than the machine-generated ones (see Appendix C for details). So, even when we 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 4c and 4d). However, as we shall see later, exploring these distances in the latent space may more appropriate and informative.

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**

5.5. Representativeness: Human Explanations are Within Distribution (the Counterfactual One)

To test the representativeness perspective, that counterfactual explanations should be within distribution and, to some degree, representative of the counterfactual class to function as useful explanations. Hence, methods with better within-distribution scores are to be preferred. But, how do the within-distribution properties of human explanations compare to those of machine explanations?

The Monte Carlo Dropout (MC-Mean, MC-Std) [40, 11] metric which measures the uncertainty in a model's prediction confidence [64, 101, 32, 11], shows that human counterfactuals are the least uncertain with respect to the model's classification [64, 40], 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[7]. In short, humans do not create visual explanations that are close to the model's decision boundary (i.e., ones with high aleatoric uncertainty [101]).

Furthermore, the R%-sub metric [100] 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, relative to class prototypes retrieved using MMD-critic [65], in contrast to the scores seen for the XAI methods. Finally, the IM1 and LOF metrics confirm this interpretation of where human counterfactuals sit, class-wise. IM1 [110], which uses an autoencoder to estimate the reconstruction error for the counterfactual class, shows that human counterfactuals have the lowest error relative to all four XAI methods for MNIST. 10-LOF [13], which is a proximity based out-of-distribution measure in the pixel space [55, 60] demonstrates that human explanations are more well grounded in the counterfactual class relative to min-edit counterfactuals.

In short, none of the current XAI methods, whether they be constraint optimisers, autoencoders or generative models [54, 110, 64, 103, 33], do a good job of corresponding to the human outputs; rather, there is a marked divergence between human and machine explanations.

5.6. Prototypicality: Human Counterfactuals Are More Prototypical

Human explanations reveal a tendency to produce counterfactuals that can be quite 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 [67, 41, 24, 122]). Figure 5 shows an image of a "helicopter"

[7]Revise failed to generate counterfactuals for several QuickDraw misclassifications, flattering it's poor performance as we considered the scenarios where it's explanations were produced in comparative experiments.

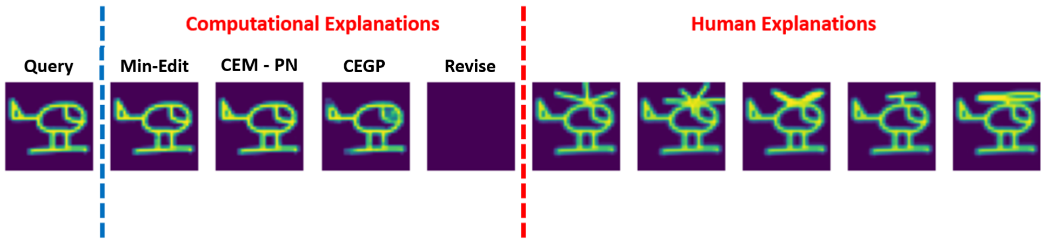


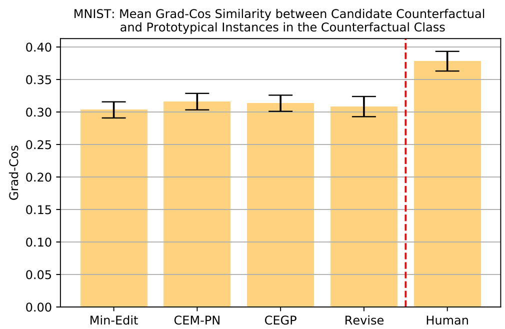
Figure 5: *Prototype Evaluations.* From the QuickDraw data, a query (a "helicopter" misclassified as a "mushroom") and the explanations generated by four XAI methods (as four counterfactual "helicopters") compared to those generated by users. Note, how people add "rotor blades", a semantic-feature, whereas the automated methods perform minimal pixel changes. Revise fails to generate an explanation, confirming findings by [53].

that was misclassified as a "mushroom", to which people add "rotor blades" to identify it as a "helicopter"[8]. In contrast, the XAI methods make small changes to a few pixels that imperceptibly modify the image.

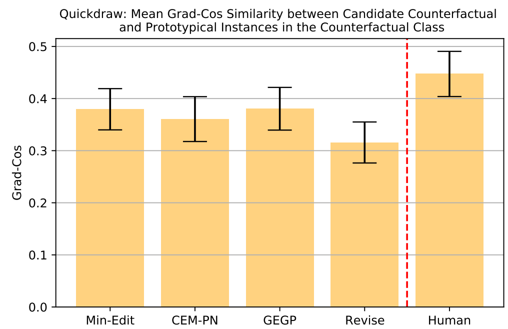
This role of prototypes in counterfactual explanations can be evaluated more directly by analysing similarities in the latent space, using the Grad-Cos similarity metric [22, 49] to compare human counterfactuals to class prototypes from the counterfactual class (retrieved using MMD-Critic [65]). As Figure 6 shows, in this latent space, human counterfactuals are more similar than all four XAI methods, to the prototypes of the counterfactual class. 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.

Humans generate counterfactual explanations that are more similar to class prototypes in the latent space, relative to other computational methods for both MNIST and QuickDraw datasets. In the case of the QuickDraw dataset, the original QuickDraw image doodles were produced under a strict time constraint of 20 seconds, which meant that some doodles were left unfinished. These unfinished, noisy doodles present significant challenges for the XAI methods, as they are degraded inputs (and possibly corrupt the ability

[8]Interestingly, in both user studies people could add pixels or erase them when producing their explanations, but they tended to add pixels more often than remove them, echoing known psychological findings. Byrne [18] noted that *"people tend to create counterfactuals about how things could have been different, that add something new to what they already know about the situation, rather than ones that remove something from it".*



(a) MNIST



(b) QuickDraw

Figure 6: *Prototype Evaluations.* 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).

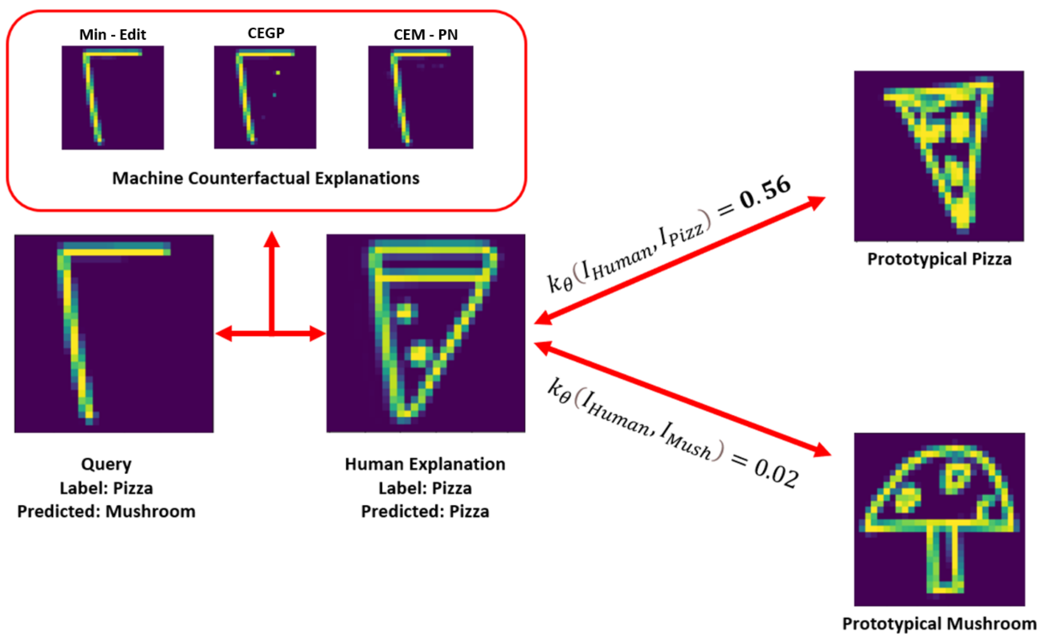


Figure 7: As the original QuickDraw images were created by humans under a 20 second time constraint, some are unfinished, presenting interpretation and reconstruction difficulties for many XAI models (see insert). In contrast, people readily fill in missing portions of the image, presumably using their knowledge of the class prototype (e.g., see similarities to MMD-Critic prototypes)

of the Revise and CEGP generative models to reconstruct the class distribution). However, people manage these degraded inputs using their knowledge of the semantic features of the represented objects. Figure 7 shows the case of an unfinished doodle of a pizza, which the model predicts to be a mushroom. All the XAI methods fail to generate plausible counterfactuals to this image, yet people simply add significant missing semantic features, leveraging their knowledge of class prototypes.

6. Conclusion: Issues, Limitations & Future Directions

Recently, many researchers have argued for a more user-centered explainable AI requiring better tests from a user perspective, a challenge that has dogged research on counterfactual XAI [89, 9, 90, 82]. 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 [79] 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 on the counterfactual world, rather than the current one. On the face of it, these results raise questions about the claimed plausibility of current methods in the literature. However, we believe that an analysis of these results with respect to "explanation-goals" yields a better interpretation of their significance. In this section, we discuss this analysis and other outstanding issues arising from the work in a discussion of the following three questions:

- How should we interpret this divergence between human- and machine-generated explanations and can this divergence be resolved?
- Can we design a prototype-driven counterfactual XAI method that better approximates human-generated counterfactuals?
- What are the limitations of the current work and what future directions could be taken to address of these limitations?

6.1. 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. How should be interpret this divergence and resolve these discrepancies? We believe the answer lies in a more thorough analysis of "explanation goals" (see Section 3.4). Conversational theories [1, 14, 15, 109] 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 [106]. This goal-based view of explanation can be used to understand the divergence between the human- and machine-generated counterfactual explanations found here. 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 different explanation goal, a "class-distribution" one, in which the counterfactual is designed to communicate broad knowledge about classes in the domain; to put it another way, people's "natural" tendency is to compute contrasting 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 critical and appropriate (e.g., in the classic recourse scenarios). However, there are also situations where distributional-explanations are critical and appropriate (e.g., in learning about domains). This analysis helps us understand the divergence between human and machine counterfactuals but is still leaves us with some issues to resolve in XAI.

First, these conclusons suggests that the community needs to co-ordinate the explanation goals being used by XAI methods with those that people naturally adopt in a given task-context. It seems highly likely that if explanation goals are not coordinated then "good" explanations will not follow, as human and machine will be literally working at cross-purposes. Indeed, an interesting empirical question for future user studies would be to determine what misunderstandings and explanatory confusions arise when explanation goals not properly coordinated.

Second, given the finding that people adopt different explanation goals, that focus on more distributional central-tendencies of classes, counterfactual XAI needs to reflect on whether min-edit-type explanations should always be the go-to option. Clearly the community needs to explore alternative counterfactual methods that better service this class-distributional goal, rather than the class-discrimination goal that is arguably over-serviced in the literature.

In the next subsection, we provide an indicative example of such a method – a prototype-driven counterfactual method – designed to better meet the class-distributional goal. A deeper issue that remains for future work is how to identify the different explanatory goals that can arise in different contexts and to determine which hold to properly deploy these alternative options for counterfactual explanation.

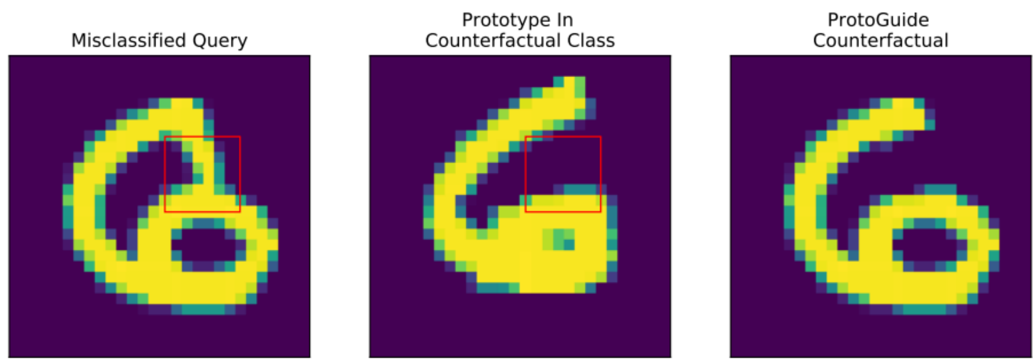
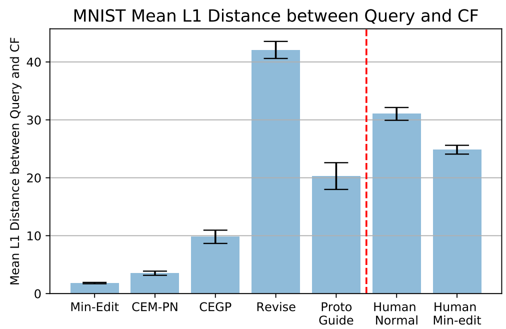


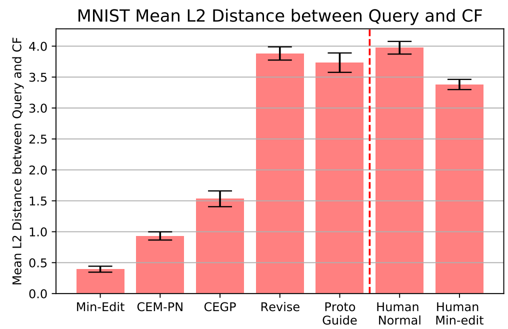
Figure 8: Given a misclassified query image, I, ProtoGuide locates the nearest prototype in the counterfactual class, I'_p, and modifies the query image with a patch from the prototype to create a candidate counterfactual explanation $I*'$.

6.2. A Prototype-Guided Counterfactual Method: ProtoGuide

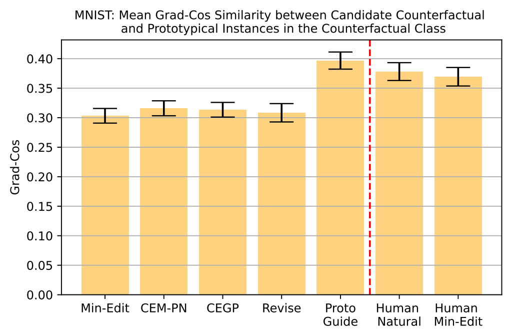
The current findings suggest we need new counterfactual methods that meet goals requiring explanations of class-distributions; recall, we saw that even though CEGP uses some prototype information, it still does not corresponds well to people's counterfactuals. To provide an indicative demonstration, we developed a novel counterfactual method that gives prototype-information a greater role in machine-generated explanations. We do not consider this method to be a definitive solution to these requirements, but advance it as a demonstration to guide future algorithmic development



(a)



(b)



(c)

Figure 9: Mean L1 (a) and L2 (b) distance scores for query-explanation pairs produced by counterfactual XAI methods including ProtoGuide (left of the dotted red line) compared to the human data (right of dotted red line), for MNIST (Error bars show standard error of the mean). Unsurprisingly, (c) ProtoGuide also produces the most similar explanations to prototypes in the latent space, mirroring the human behaviour according to the Grad-Cos similarity metric.

Guided by evidence from the human counterfactual explanations, the proposed method, ProtoGuide, leverages information from class prototypes retrieved using MMD-critic [65] and can be cast as an endogenous technique [46]. Given a query image, I, a similar prototype, I'_p, from the counterfactual class is retrieved (using e.g., cosine similarity) such that $b(I) = c$ and $b(I'_p) = c'$. Next, a counterfactual $I*'$ is constructed in the feature space by replacing spatial cells from $f(I)$ with cells from $f(I'_p)$. The size of the modified cell patch is a parameter that we set to a default value of 7×7. Following [44, 114], a permutation matrix P selectively replaces entries in $f(I)$, aiming to avoid the trivial solution of replacing all cells in $f(I)$ with the cells from $f(I'_p)$. As there are many possible modifications that could create a counterfactual, the swap that maximises the structural similarity between I'_p and $I*'$ is selected.

We have implemented the ProtoGuide method and applied it on the MNIST misclassifications used in our original study, with the corresponding CNN as the underlying black-box model. We then tested ProtoGuide's counterfactual explanations using our evaluation metrics to compare them to the human explanations. These evaluations show that ProtoGuide better approximates human-generated explanations, reducing the divergence previously reported between human- and machine-generated counterfactuals (see Figures 8 and 9). As such, ProtoGuide demonstrates the feasibility of a new family of counterfactual algorithms that can service class-distribution goals when domain learning is required as opposed to class-discriminative goals when, for example, minimal, algorithmic recourse is required.

6.3. Limitations & Future Directions

Although the present work establishes several new findings for the XAI literature, it does have limitations that deserve attention in future work relating to (i) explanation-goal analyses of XAI scenarios, (ii) the performance of further psychological tests, (iii) the development of prototype-driven methods for counterfactual XAI, and (iv) the datasets used.

Explanation Goals. Sørmo et al. [106] argued that XAI should be cast as conversational interactions between agents that fundamentally rely on the explanation goals of those agents (for related arguments in psychology see [51, 86]). They cite Achinstein's [1] view of explanation as an "illocutionary act", an "explaining act (that) defines some aspect of the context and purpose behind the explanation, which is needed for a correct and meaningful interpretation of the explanation product" (Sørmo et al., 2005, pp. 114-115). Arguably, counterfactual XAI has not given sufficient attention to this aspect of explanation scenarios. We have seen that most counterfactual methods implicitly assume that a class-discrimination goal for explanation, but people assume a class-distribution goal. This finding suggests that much more work needs to be done, both psychologically and computationally, for this very different explanation scenario. Indeed, apart from developing methods that can meet class-distribution goals, we also might need to consider methods that can handle diverse explanation goals in some unitary mechanism that automatically selects the goal-appropriate counterfactuals to be used in a given situation. The present work only begins to scratch the surface of such questions and issues.

Psychological Testing. Though the present work focuses centrally on the user testing of counterfactual explanation, more work needs to be done to follow up on its findings. The MNIST study reported here was carefully designed using an appropriate power analysis to test people's generation of explanations. Our second study, involving QuickDraw, was a pilot experiment and, as such, strictly-speaking may need to be re-run with a larger sample. However, given the regularity of people's individual responses to the QuickDraw images and the close correspondence between these results and those seen in the MNIST study, we would expect the QuickDraw findings to be replicated. Furthermore, it might also make sense to develop testing methods that can be run on crowd-sourced platforms, rather than continuing to use face-to-face methods, given their overhead in time and effort. However, such a move is not as straightforward as it might seem. There would be challenges running the drawing tool on such platforms as they tend to use static selection interfaces. This means that one would have to present participants with static options for the counterfactual explanations, perhaps based on people's responses from the current studies. This crowd-sourcing option would facilitate more user testing but would lack the open-ended nature of the current tests. So, there are pros and cons to be assessed with respect to what can be discovered using these alternative experimental paradigms.

Algorithmic Development. The user study findings reported here highlight the need for further exploration of prototype-driven counterfactual methods. In the previous sub-section, we described an initial, novel design for such a method and, briefly, showed how it produced outputs much closer to human-generated explanations. However, this proposed method is really more a demonstration than a definitive proposal on how prototype-driven counterfactuals might be computed. Considerably more needs to be done by the community to determine the best prototypes to compute and the most appropriate ways to deploy such methods in different explanation scenarios. One promising avenue is to leverage self-supervised learning models to learn semantically meaningful and discriminative regions to modify when creating counterfactuals (e.g., the rotor in an image of a helicopter) [111, 121].

More Complex Datasets. Obviously, many other datasets could and should be tested beyond the two considered here. Note, that 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). Gathering human-generated explanations for photo-real objects would be very interesting though further tool-development and thought would be required to provide a suitable editing tool (e.g., an editing tool that simply removes or adds pixels might not result in realistic or meaningful explanations). One promising line of work for this would be to use text annotations from users in combination with recently developed generative models [95] to create realistic counterfactual explanations. Indeed, these considerations underscore the challenges in performing user-centered studies, perhaps indicating why they have been so rare in XAI literature to date.

6.4. Closing Comments

The present work promotes a user-centered approach to counterfactual XAI, evaluating the differences between explanations that are generated by people and machines using popular benchmark comparison metrics from the counterfactual literature (e.g., distance, representativeness and latent similarity). 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.

7. Data and Code Availability

The MNIST dataset is openly available[9] and the Google QuickDraw dataset is made available by Google, Inc. under the Creative Commons Attribution 4.0 International license[10]. All data from the user study is made available under a MIT licence[11]. The code used to detail hyperparameters and produce our results is available under the MIT Licence[12].

[9] http://yann.lecun.com/exdb/mnist/

[10] https://github.com/googlecreativelab/quickdraw-dataset

[11] https://github.com/e-delaney/cfe_images_how_people_differ_from_machines/tree/main/User_data

[12] https://github.com/e-delaney/cfe_images_how_people_differ_from_machines

8. 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). Thanks also to Greta Warren for advice on the statistical issues, the relevant literature and psychological measures of counterfactual explanation.

Appendix A. Monte Carlo Dropout Technical Details

Following the description of [30, 32], we provide an overview of how MC-Dropout can be applied. When a predictive distribution $p(y|x, D)$ is obtained, the corresponding uncertainty can be uncovered by analysing the variance.

From a Bayesian perspective, total predictive uncertainty, $\mathbb{V}(y \mid x)$, can be decomposed into a sum of two components, namely epistemic (model) uncertainty and aleatoric (data) uncertainty [62, 30]. Let x represent some input, y the target variable and Θ the random parameters of the model, then

$$\mathbb{V}(y \mid x) = \underbrace{\mathbb{V}(\mathbb{E}(y \mid x, \Theta))}_{Epistemic} + \underbrace{\mathbb{E}(\mathbb{V}(y \mid x, \Theta))}_{Aleatoric} \tag{A.1}$$

To learn this distribution we learn the distribution over the model parameters p($\Theta|D$) (i.e., the parametric posterior distribution).

Work in [40] showed that, by randomly switching off neurons in a neural network using different dropout configurations, one could approximate the parametric posterior distribution without the need to retrain the network. Each dropout configuration Θ_t corresponds to a sample from the approximate parametric posterior distribution $q(\Theta|D)$ s.t. $\Theta_t \sim q(\Theta|D)$.

Sampling from the approximate posterior enables us to uncover the predictive distribution $p(y \mid x)$:

$$p(y \mid x, D) \approx \int_{\Omega} \underbrace{p(y \mid x, \Theta)}_{likelihood} \underbrace{q(\Theta \mid D)}_{posterior} \, d\Theta \tag{A.2}$$

The above integral can be approximated through Monte Carlo methods, giving;

$$p(y \mid x, D) \underbrace{\approx}_{MC} \frac{1}{T} \sum_{t=1}^{T} p(y \mid x, \Theta_t) \tag{A.3}$$

Multiple forward passes with different dropout configurations allow one to uncover the predictive distribution. Under the assumption that the likelihood is Gaussian distributed, the mean $f(x, \theta)$ and variance $s^2(x, \Theta)$ parameters of the Gaussian function are determined from Monte Carlo simulation such that $f(x, \theta), s^2(x, \Theta) \sim$ MC-Dropout(x), and can yield useful information about the predictive uncertainty through connecting back with Equation A.2.

Appendix B. Prototype Evaluations: Prototypes & Similarity

To determine the closeness of generated counterfactuals to the prototype(s) of the counterfactual class, MMD-Critic [65] was used to generate 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. [65], 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 [91, 65], 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:

$$MMD^2 = \frac{1}{m^2} \sum_{i,j=1}^{m} k(z_i, z_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(z_i, x_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, x_j) \tag{B.1}$$

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:

$$k(x, x') = exp\,(-\gamma\| |x - x'\| |_2) \tag{B.2}$$

The MMD^2 measure, kernel function and greedy search are combined in an algorithm to find prototypes [91]. 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.

Appendix C. Additional User Study Details and Results

participants who were instructed to minimally edit the query to generate a counterfactual explanation (Min-Edit group) made significantly smaller edits, measured by L1 and L2 distance between the query and generated counterfactual, than the group not given these instructions (Normal group); tested using a two-sample t-test, $p < 0.001$) (See Figure C.10).

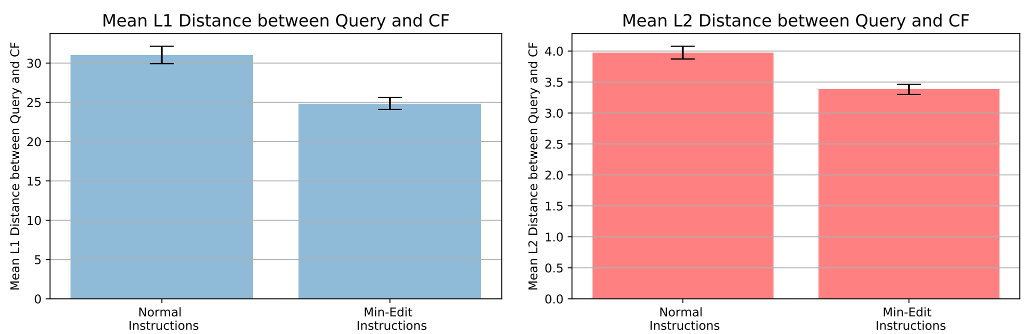


Figure C.10: *Proximity Evaluations.* Mean L1 and L2 distances between the misclassified query and the corresponding counterfactual. Explanations produced by the two human groups (Normal & Min-Edit groups) for the MNIST dataset (error bars show standard error of the mean).

References

[1] P. Achinstein, The nature of explanation, Oxford University Press, 1983.

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

[3] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems, 9505–9515, 2018.

[4] A. Akula, S. Wang, S.-C. Zhu, Cocox: Generating conceptual and counterfactual explanations via fault-lines, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2594–2601, 2020.

[5] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, 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, International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088, 2019.

[6] A. Artelt, B. Hammer, On the computation of counterfactual explanations–A survey, arXiv preprint arXiv:1911.07749 .

[7] S. Aryal, M. T. Keane, "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.

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

[9] S. Barocas, A. D. Selbst, M. Raghavan, The hidden assumptions behind counterfactual explanations and principal reasons, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 80–89, 2020.

[10] A. Bäuerle, H. Neumann, T. Ropinski, Training De-Confusion: An Interactive, Network-Supported Visual Analysis System for Resolving Errors in Image Classification Training Data, arXiv preprint arXiv:1808.03114 .

[11] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, 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, 401–413, 2021.

[12] A. Birhane, V. U. Prabhu, J. Whaley, Auditing Saliency Cropping Algorithms, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 4051–4059, 2022.

[13] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, LOF: identifying density-based local outliers, in: ACM SIGMOD, 93–104, 2000.

[14] S. Bromberger, An approach to explanation, Analytical philosophy 2 (1965) 72–105.

[15] S. Bromberger, Why-questions, in: R. Colodny (Ed.), Mind & Cosmos, vol. 42, Pittsburg University Press, Pittsburg, USA, 86–111, 1966.

[16] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 454–464, 2020.

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

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

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

[20] L. Celar, R. M. Byrne, How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains, Memory & Cognition (2023) 1–16.

[21] C.-H. Chang, E. Creager, A. Goldenberg, D. Duvenaud, Explaining image classifiers by counterfactual generation, arXiv preprint arXiv:1807.08024 .

[22] G. Charpiat, N. Girard, L. Felardos, Y. Tarabalka, Input similarity from the neural network perspective, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 5342–5351, 2019.

[23] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. K. Su, This looks like that: deep learning for interpretable image recognition, in: Advances in Neural Information Processing Systems, vol. 32, 2019.

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

[25] L. Connell, M. T. Keane, What plausibly affects plausibility? Concept coherence and distributional word coherence as factors influencing plausibility judgments, Memory & cognition 32 (2) (2004) 185–197.

[26] L. Connell, M. T. Keane, A model of plausibility, Cognitive Science 30 (1) (2006) 95–120.

[27] F. J. Costello, M. T. Keane, Efficient creativity: Constraint-guided conceptual combination, Cognitive Science 24 (2) (2000) 299–349.

[28] X. Dai, M. T. Keane, L. Shalloo, E. Ruelle, R. M. Byrne, Counterfactual explanations for prediction and diagnosis in xai, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 215–226, 2022.

[29] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, in: Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I, Springer, 448–469, 2020.

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

[31] E. Delaney, D. Greene, M. T. Keane, Instance-based counterfactual explanations for time series classification, in: International Conference on Case-Based Reasoning, Springer, 32–47, 2021.

[32] E. Delaney, D. Greene, M. T. Keane, Uncertainty Estimation and Out-of-Distribution Detection for Counterfactual Explanations: Pitfalls and Solutions, arXiv preprint arXiv:2107.09734 .

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

[34] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, C. Dugan, Explaining models: an empirical study of how explanations impact fairness judgment, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 275–285, 2019.

[35] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 .

[36] E. Erdfelder, F. Faul, A. Buchner, GPOWER: A general power analysis program, Behavior research methods, instruments, & computers 28 (1) (1996) 1–11.

[37] C. Ford, M. Keane, Explaining Classifications to Non-Experts: An XAI User Study of Post-Hoc Explanations for a Classifier When People Lack Expertise, in: 26th International Conference on Pattern Recognition, 2022.

[38] M. Förster, P. Hühn, M. Klier, K. Kluge, Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 1274, 2021.

[39] M. Förster, M. Klier, K. Kluge, I. Sigler, Evaluating explainable Artifical intelligence–What users really appreciate .

[40] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, 1050–1059, 2016.

[41] A. Ghorbani, J. Wexler, J. Y. Zou, B. Kim, Towards automatic concept-based explanations, Advances in Neural Information Processing Systems 32.

[42] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, in: International Conference on Learning Representations, 2015.

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

[44] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: ICML, PMLR, 2376–2384, 2019.

[45] R. M. Grath, L. Costabello, C. L. Van, P. Sweeney, F. Kamiab, Z. Shen, F. Lecue, Interpretable credit application predictions with counterfactual explanations, arXiv preprint arXiv:1811.05245 .

[46] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55.

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

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

[49] K. Hanawa, S. Yokoi, S. Hara, K. Inui, Evaluation of Similarity-based Explanations, in: International Conference on Learning Representations, 2021.

[50] L. A. Hendricks, R. Hu, T. Darrell, Z. Akata, Generating counterfactual explanations with natural language, arXiv preprint arXiv:1806.09809 .

[51] D. J. Hilton, Conversational processes and causal explanation., Psychological Bulletin 107 (1) (1990) 65.

[52] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: Challenges and prospects, arXiv preprint arXiv:1812.04608 .

[53] B. Höltgen, L. Schut, J. M. Brauner, Y. Gal, DeDUCE: Generating Counterfactual Explanations Efficiently, arXiv preprint arXiv:2111.15639 .

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

[55] K. Kanamori, T. Takagi, K. Kobayashi, H. Arimura, DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization, in: IJCAI-20, 2855–2862, 2020.

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

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

[58] D. Kaushik, E. Hovy, Z. Lipton, Learning The Difference That Makes A Difference With Counterfactually-Augmented Data, in: International Conference on Learning Representations, URL `https://openreview.net/forum?id=Sklgs0NFvr`, 2020.

[59] M. T. Keane, E. M. Kenny, 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), Springer, 155–171, 2019.

[60] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21), 2021.

[61] M. T. Keane, B. Smyth, Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI), in: ICCBR, Springer, 163–178, 2020.

[62] A. Kendall, Y. Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

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

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

[65] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: Advances in Neural Information Processing Systems, 2280–2288, 2016.

[66] B. Kim, C. Rudin, J. A. Shah, The bayesian case model: A generative approach for case-based reasoning and prototype classification, in: Advances in neural information processing systems, 1952–1960, 2014.

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

[68] L. Kirfel, A. Liefgreen, What If (and How...)? - Actionability Shapes People's Perceptions of Counterfactual Explanations in Automated Decision-Making, in: ICML Workshop on Algorithmic Recourse, 2021.

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

[70] G. Klein, M. Jalaeian, R. Hoffman, S. Mueller, The Plausibility Gap: A model of sensemaking .

[71] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: International conference on machine learning, PMLR, 1885–1894, 2017.

[72] U. Kuhl, A. Artelt, B. Hammer, Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting, arXiv preprint arXiv:2205.05515 .

[73] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, F. Doshi-Velez, An evaluation of the human-interpretability of explanation, arXiv preprint arXiv:1902.00006 .

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

[75] R. Larasati, A. De Liddo, E. Motta, 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.

[76] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, in: IJCAI-19, 2801–2807, 2019.

[77] M. L. Leavitt, A. Morcos, Towards falsifiable interpretability research, in: NeurIPS 2020 Workshop on ML Retrospectives, Surveys & Meta Analyses, 2020.

[78] Y. LeCun, The MNIST database of handwritten digits, http://yann. lecun. com/exdb/mnist/ .

[79] D. Lewis, Counterfactuals, John Wiley & Sons, 2013.

[80] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[81] B. Y. Lim, A. K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2119–2128, 2009.

[82] B. Y. Lim, Q. Yang, A. M. Abdul, D. Wang, Why these explanations? Selecting intelligibility types for explanation goals., in: IUI Workshops, 2019.

[83] Z. C. Lipton, The mythos of model interpretability, arXiv preprint arXiv:1606.03490 .

[84] A. Lucic, H. Haned, M. de Rijke, Why does my model fail? contrastive local explanations for retail forecasting, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 90–98, 2020.

[85] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 4765–4774, 2017.

[86] B. F. Malle, How the mind explains behavior, MIT press, 2006.

[87] D. Martens, F. Provost, Explaining data-driven document classifications, MIS quarterly 38 (1) (2014) 73–100.

[88] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling, in: 2021 IEEE Symposium on Computers and Communications (ISCC), IEEE, 1–7, 2021.

[89] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[90] T. Miller, Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support, arXiv preprint arXiv:2302.12389 .

[91] C. Molnar, Interpretable machine learning, Lulu.com, 2020.

[92] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 607–617, 2020.

[93] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, G. Klein, 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 .

[94] C. Nugent, P. Cunningham, A case-based explanation system for black-box systems, Artificial Intelligence Review 24 (2) (2005) 163–178.

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

[96] M. T. Ribeiro, S. Singh, C. Guestrin, 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, ACM, 1135–1144, 2016.

[97] A. Ross, F. Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

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

[99] C. Russell, Efficient search for diverse coherent explanations, in: Conference on Fairness, Accountability, and Transparency, 20–28, 2019.

[100] P. Samangouei, A. Saeedi, L. Nakagawa, N. Silberman, Explaingan: Model explanation via decision boundary crossing transformations, in: European Conference on Computer Vision (ECCV), 666–681, 2018.

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

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

[103] S. Singla, B. Pollack, J. Chen, K. Batmanghelich, Explanation by Progressive Exaggeration, in: International Conference on Learning Representations, 2020.

[104] B. Smyth, M. T. Keane, A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations, in: International Conference on Case-Based Reasoning, Springer, 18–32, 2022.

[105] K. Sokol, P. A. Flach, Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety, in: SafeAI@ AAAI, 2019.

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

[107] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.

[108] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: A comparison of rule-based and example-based explanations, Artificial Intelligence 291 (2021) 103404.

[109] B. C. Van Fraassen, et al., The scientific image, Oxford University Press, 1980.

[110] A. Van Looveren, J. Klaise, 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, Springer, 650–665, 2021.

[111] S. Vandenhende, D. Mahajan, F. Radenovic, D. Ghadiyaram, Making heads or tails: Towards semantically consistent visual counterfactuals, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII, Springer, 261–279, 2022.

[112] S. Verma, J. Dickerson, K. Hines, Counterfactual Explanations for Machine Learning: A Review, arXiv preprint arXiv:2010.10596 .

[113] S. Verma, J. Dickerson, K. Hines, Counterfactual explanations for machine learning: Challenges revisited, arXiv preprint arXiv:2106.07756 .

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

[115] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, Harv.J.Law Tech. 31 (2017) 841.

[116] G. Warren, R. M. J. Byrne, M. T. Keane, Categorical and Continuous Features in Counterfactual Explanations of AI Systems, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23, Association for Computing Machinery, New York, NY, USA, 171–187, 2023.

[117] G. Warren, M. T. Keane, R. M. J. Byrne, Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI, in: IJCAI-ECAI'22 Workshop: Cognitive Aspects of Knowledge Representation, 2022.

[118] G. Warren, B. Smyth, M. T. Keane, "Better" Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI), in: International Conference on Case-Based Reasoning, Springer, 63–78, 2022.

[119] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (1) (2019) 56–65.

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

[121] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 3733–3742, 2018.

[122] R. Zhang, P. Madumal, T. Miller, K. A. Ehinger, B. I. Rubinstein, Invertible concept-based explanations for cnn models with non-negative concept activation vectors, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 11682–11690, 2021.

[123] W. Zhao, S. Oyama, M. Kurihara, Generating natural counterfactual visual explanations, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 5204–5205, 2021.

[124] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2921–2929, 2016.