

If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques

Mark T. Keane^{1,2,3}, Eoin M. Kenny^{1,3}, Eoin Delaney^{1,3} and Barry Smyth^{1,2}

¹School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

²Insight Centre for Data Analytics, Belfield, Dublin 4, Ireland

³VistaMilk SFI Research Centre, Belfield, Dublin 4, Ireland

{mark.keane, barry.smyth}@ucd.ie, {eoin.kenny1, eoin.delaney4}@ucdconnect.ie

Abstract

In recent years, there has been an explosion of AI research on counterfactual explanations as a solution to the problem of eXplainable AI (XAI). These explanations seem to offer technical, psychological and legal benefits over other explanation techniques. We survey 100 distinct counterfactual explanation methods reported in the literature. This survey addresses the extent to which these methods have been adequately evaluated, both psychologically and computationally, and quantifies the shortfalls occurring (e.g., only 21% of these methods have been user tested). Five key deficits in the evaluation of these methods are detailed and a roadmap, with standardized benchmark evaluations, is proposed to resolve the issues arising; issues, that currently block scientific progress in this field.

1 Introduction

In recent years, as AI systems are increasingly deployed in everyday life, there has been a slew of research papers on the problem of eXplainable AI (XAI), driven by concerns about whether these systems are fair, accountable and trustworthy. In the literature on post-hoc explanations, that aim to justify an AI model's predictions *after the fact*, the use of counterfactual explanations has gained considerable traction based on claimed technical, psychological and legal benefits (see Figure 1). Counterfactual explanations provide information to users on what might be done to *change* the outcome of an automated decision (e.g., “if your paper had more novelty, it would have been accepted to this conference”). In this paper, we critically review the evaluations carried out on counterfactual explanation methods, focusing on psychological issues. To put it simply, we assess whether there is any evidence that counterfactuals explanations “work” and whether the properties promulgated in current methods are relevant to end users. So, this review

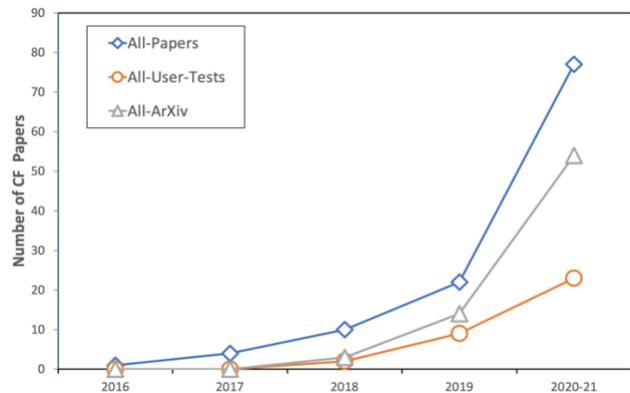


Figure 1: The number of surveyed papers (per annum) on counterfactual XAI (2016-2021) on (a) CF methods (*All-Papers-blue*) (b) CF XAI user studies (*All-User-Tests - orange*) (c) ArXiv from a search of abstracts using the terms “counterfactual explanation” (*All-ArXiv- grey*)¹ [n.b., only a subset is cited here¹]

is, in part, a critique on the paucity of user testing in the area, but it also tries to go beyond this critique to formulate a firmer basis for better evaluation of counterfactual methods in the future (henceforth, we abbreviate “counterfactual” as CF).

Obviously, this analysis is also relevant to the bigger picture for XAI, where many have pointed to its failure to properly address user requirements [Anjomshoae *et al.*, 2019; Hoffman and Zhao, 2020]. As in other areas of XAI, CF XAI also suffers from an “over-reliance on intuition-based approaches” that, arguably, impedes scientific progress [Leavitt and Morcos, 2020]. Our main concern is that the XAI community is busily developing technical solutions that may have no practical benefits to people in real-life. As Barocas *et al.* [2020] have pointed out, on CF explanations, there is a disconnect between the features of a model and actions in the real-world, that needs to be bridged for these explanation techniques to work successfully.

To be positive, we do not just diagnose the problem but advance solutions in the form of a roadmap for future evaluations, one that argues for standardized, benchmark metrics that are psychologically grounded. Our hope is that these proposals for CF explanation, will also provide a more general template for how XAI might approach user-requirement issues and advance scientific progress.

In the remainder of this introduction, we briefly consider why counterfactuals have attracted so much attention in XAI, before summarizing the contributions of the current paper. Then, we review the key insights underlying current CF XAI research (see section 2). In the remainder of the paper, we outline five key deficits in CF research that need to be remedied if progress is to be made, focused largely on the issue on the psychological validity of current proposals. As we shall see, CF research like many areas of XAI, claims that its intuition-based computational evaluations are proxies for psychological evaluations when such claims are at best, unsubstantiated, and, at worst, just wrong.

1.1 Why Counterfactuals?

Intuitively, counterfactual explanations seem to convey a lot more information about a decision, prediction, or classification than factual explanations [Keane and Kenny, 2021]. To re-run the classic example, a loan refusal by an automated AI system using a *factual explanation* might explain the decision to a customer by saying “you did not get the loan because your profile is similar to applicants who were also refused” [Kenny and Keane, 2019; Kenny *et al.*, 2021]. In contrast, a *counterfactual explanation* might explain the same decision by saying “if you earned \$1k more, then you would have gotten the loan”. Not only does this explanation inform end-users about key features leading to the decision, but it also opens the door to allowing them to improve their chances of success; that is, the explanation can identify “actionable” features that permit a remediation of the decision (i.e., algorithmic recourse). Proponents of CFs argue for their significant technical, psychological and legal benefits over other explanation techniques.

Technically, CF techniques build on a wave of popular XAI work showing that feature-importance analyses of ML models can provide automated explanations. Techniques – such as LIME and SHAP – showed that presenting information on the relative importance of features for a model’s predictions can provide acceptable automated explanations [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017]. Some CF methods carry on directly from this approach in selecting features to change in the CF [Pedreschi *et al.*, 2019]. Wacheter *et al.*’s, [2018] seminal work casts it as an optimisation problem balancing the proximity of the CF to the test-instance and its proximity to the decision boundary. These works demonstrated the technical feasibility of finding automated explanations that appeared to be plausible to users.

Psychologically, this computational work has been backed by a long-standing literature in Psychology [Byrne, 2007, 2019; Mueller *et al.*, 2019] and Philosophy [Lewis,

1973/2013; Woodward, 2005] arguing for the centrality of counterfactuals in human cognition, explanation and science. But, it was Miller’s [2019] seminal review of this literature that perhaps convinced the XAI community of the importance of these explanations. These works solidified the view that CFs could provide psychologically intuitive, plausible explanations; especially, if they were *sparse* (had few feature differences) and *proximate* (the *closest possible world*). However, as we shall see, solid evidence for many of these claims has not always been forthcoming.

Legally, the final fillip to the counterfactual program came from the argument that CF explanations were compliant with the EU’s General Data Protection Regulation (GDPR) [Wacheter *et al.*, 2018]; so, CFs seemed to provide an explanation method that meets emerging regulatory requirements on AI systems. However, note, that this claim rests on the assumption that people find CF explanations comprehensible, which remains to be proven sufficiently.

1.2 Motivation & Novel Contributions

In 2020, two reviews of the CF XAI literature, between them, surveyed 52 CF techniques [Karimi *et al.*, 2020a; Verma *et al.*, 2020]. However, given the breakneck pace of this area, we have uncovered a further 48 CF-method papers not referenced in these reviews (see Figure 1). Furthermore, these surveys do not focus on user studies, but rather on the technical features of explanatory methods. Thus, the present survey complements these earlier surveys, but offers a very different analysis of the area focusing on the key insights made, user studies on CF XAI, and the evaluation deficits in the field. As such the key contributions of the paper are:

- A novel, critical (updated) review of 100 CF XAI methods, focused on -- psychological (user studies) and computational (metrics) -- evaluation deficits
- The quantification of these deficits, to show the extent of underperformance in the research effort, to-date
- The proposal of a roadmap to standardize future evaluations and place them on a firmer psychological footing for scientific progress to be achieved.

As we shall see, the current critique rests on identifying the paucity of user studies on CF XAI (see Figure 1). This neglect undermines the “proxy” computational evaluations currently used, as they are intuition-based and not adequately ground-truthed. In the next section, we summarize the main “insights” driving CF XAI research (section 2). Then, with this analysis as a guide, we detail five main deficits in the area covering (i) the lack of user studies (section 3), (ii) the definition of plausibility (section 4), (iii) the issue of sparsity (section 5), (iv) the assessment of coverage (section 6) and (v) comparative testing (section 7). We conclude with proposals on how to rectify these deficits in future CF XAI work with a roadmap and benchmark evaluation metrics (section 8).

2 Counterfactual Insights

The previous reviews of CF XAI mainly profile techniques in terms of their technical properties. We offer a different analysis based on the high-level “insights” that have inspired the field. We use this discovery-based approach as it forms the basis for our proposed standardized evaluation metrics; specifically, that key evaluation metrics should address the key “insights” underlying the field (see section 8). Although, there is huge diversity in the 100 distinct CF methods surveyed here, there are perhaps four big ideas that have driven the area forward: namely, that generated CF explanations need to be (i) guided by proximity, (ii) feature focused, (iii) distributionally-faithful and, possibly, (iv) instance-based. Not all methods sign-up to these insights but many current proposals are designed to meet them as high-level requirements. They are summarized below.

CFs are proximity-guided. As we have seen, the seminal work on CF explanation [Wacheter *et al.*, 2018; Mittelstadt *et al.*, 2019] proposes perturbing the features of synthetic CF instances, under a loss function balancing proximity to the test-instance against proximity to the decision boundary for the CF class, using a scaled L_1 -norm distance-metric. This idea has inspired follow-on work using different distance metrics (e.g., L_2 -norm) or, indeed, combinations of distance metrics [Dandl *et al.*, 2020; Artelt and Hammer, 2020], with added constraints to deliver *diverse* CFs [Mothilal *et al.*, 2020; Russell, 2019]. Hence, later, we will argue for the use of selected distance metrics to benchmark evaluations (ideally, ones that are psychologically grounded).

CFs need to focus on features. The second major insight that quickly emerged in the area, was on the need to focus on the “right” features (actionable ones) to perturb and avoid the “wrong” ones (immutable ones), by using *predictive importance* [Maartens and Provost, 2014; Guidotti *et al.*, 2018; Pedreschi *et al.*, 2019], “actionability” [Ustun *et al.*, 2019; Karimi *et al.*, 2020b; Chou *et al.*, 2021], “coherence” [Russell, 2019; Gomez *et al.*, 2020] and “causality” [Karimi *et al.*, 2020c; Chou *et al.*, 2021]; while also considering dependencies between features [Mothilal *et al.*, 2020; Kanamori *et al.*, 2020]. In our evaluation guidelines we discuss some of the issues around evaluating this insight.

CFs need to be distributionally-faithful. It also became clear that the CFs need to fall within the dataset’s distribution to ensure plausibility, as some methods produced out-of-distribution, invalid data-points [Wachter *et al.*, 2018; Laugel *et al.*, 2019] (see Figure 2). Hence, methods emerged using generative models [Dhurandhar *et al.*, 2018; Joshi *et al.*, 2019; Liu *et al.*, 2019; Singla *et al.*, 2019] and/or techniques manipulating the latent features of models [Hendricks *et al.*, 2018; Van Looveren and Klaise, 2019; Akula *et al.*, 2020; Kenny and Keane, 2021]. So, benchmark metrics need to assess the distributional properties of CFs and their coverage.

CFs are instance-guided: Finally, some have argued that the best way to generate “good” CFs is to rely on the dataset, either (a) directly by using Nearest Unlike Neighbours (NUNs) [Nugent *et al.*, 2009] or (b) indirectly, by adapting instances [Goyal *et al.*, 2019; Keane and Smyth, 2020;

Delaney *et al.*, 2020; Smyth and Keane, 2021]; closely related methods preferentially select synthetic CFs falling in dense regions of the dataset [Poyiadzi *et al.*, 2020]. We propose a relative-distance metric to capture this.

In the following five sections, we critique the psychological and computational evaluations used in CF XAI, before considering how the deficits found might be rectified.

3 Deficit #1: Neglecting Users

The neglect of user studies is the “original sin” of XAI research. In an XAI-wide survey, Adadi and Berrada [2018] reported that only 5% of papers evaluated interpretability. In the CF XAI papers surveyed here, only 31% of papers perform user studies (36 out of 117) and fewer (21%) directly user-test a specific CF method (i.e., many are non-model tests). Indeed, as the area expands exponentially, the relative proportion of user studies is decreasing (see Figure 1). Furthermore, many of these studies are methodologically questionable (e.g., use low Ns, poor or inappropriate statistics, unreproducible designs). Many user studies test the use of CFs as explanations relative to no-explanation controls, rather than specifically testing a given method. So, to use a blunt metaphor, we may be fine-tuning AI methods with elaborate bells and whistles that no human-ear can hear.

To be more positive, most studies reporting these tests do show CFs to be useful and sometimes preferred by end users (e.g., [Lim *et al.*, 2009; Dodge *et al.*, 2019]). Lim *et al.* [2009] tested the use of What-if, Why-Not, How-to and Why explanations and found that they all improved performance relative to no-explanation controls. Dodge *et al.* [2019] assessed four different explanation strategies (e.g., case based, counterfactual, factual) on biased/unbiased classifiers and found counterfactual explanations to be the most impactful. However, it should also be said, that some studies show that CF explanations often require greater cognitive effort and do not always outperform other methods [Lim *et al.*, 2009; Lage *et al.*, 2019; van der Waa *et al.*, 2021].

Notably, however, few of these studies directly test a particular CF method. We have found that only 25 papers (out of 100 method papers) perform user studies on the proposed CF method (e.g., [Goyal *et al.*, 2019; Singla *et al.*, 2019; Lucic *et al.*, 2020]). These studies typically show some improvement in people’s performance on or judgement of an AI system, relative to no-explanation controls. Finally, even fewer of these studies pit one method against another in a comparative user study (only 7 out of 100 method-papers; e.g., [Akula, *et al.*, 2020; Förster *et al.*, 2020a, 2020b]).

In summary, this means that from 100 distinct CF algorithms in the literature, only 7% report user studies that specifically address the detailed properties of their methods. This state-of-affairs has knock-on effects for other aspects of CF evaluation. Basically, we do not know whether one method is better than another on (i) plausibility (section 4), (ii) sparsity (section 5), or (iii) coverage of representative problems (section 6). In the following sections, we detail these deficits in the evaluation of CF techniques.

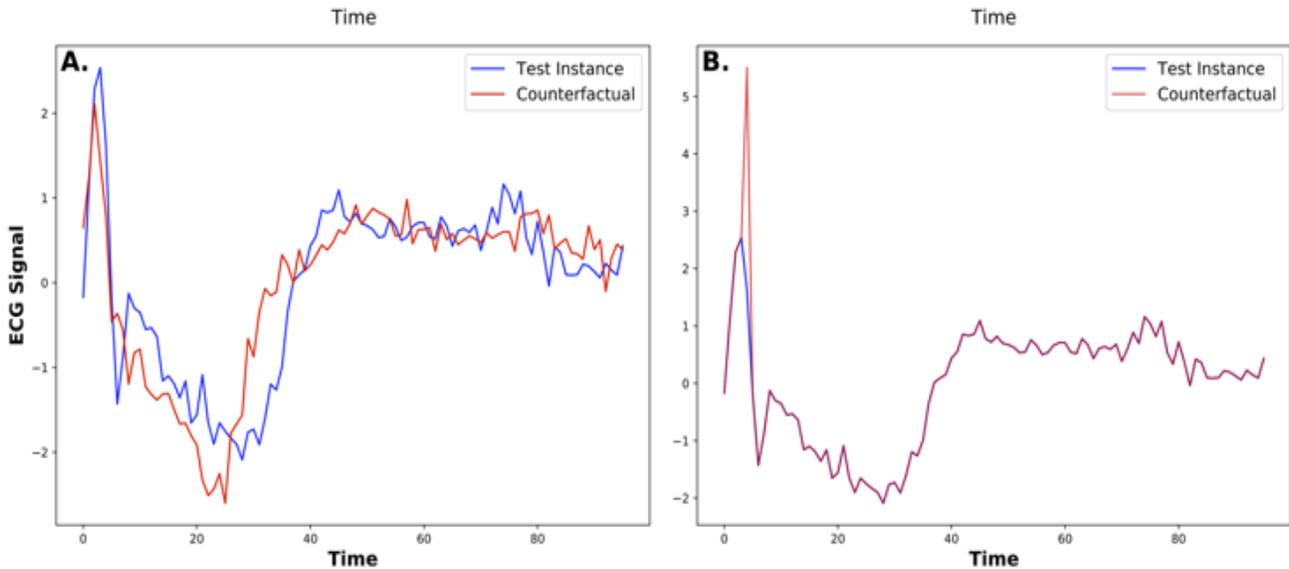


Figure 2: Two counterfactual comparisons for a time-series EEG signal (A) a “native CF” (a NUN from the dataset) that is distant from the test instance but within distribution (L_1 -norm = 38.29) and (B) a synthetic CF generated by a proximity method [Wachter *et al.*, 2018] which is much closer to the test instance (L_1 -norm = 3.88) and offers a sparser solution but is out-of-distribution according to the OCSVM metric (given the spike). Neither of these counterfactuals are likely to be plausible to end-users (from [Delaney *et al.*, 2020]).

4 Deficit #2: What’s Plausible?

All counterfactual explanation methods make the core claim that their method delivers *plausible* or *relevant* or *feasible* or *helpful* or *good* explanations to human users, explanations that meet user/stakeholder goals in diverse AI systems. Different techniques define plausibility in different ways, depending on the insights they promote (see section 2): so “plausible” CFs are (i) close to the test instances and the decision boundary, (ii) use the “right” features (e.g., mutable, actional, causal), (iii) faithful to the distribution, and/or (iv) from the training data. For the most part, these claims are not supported by user tests but are declared on intuitive grounds. Support for these claims is provided by “proxy” computational evaluations that often directly parallel the theoretical claim (e.g., a generative technique is plausible because it produces within-distribution CFs)¹. However, all of these claims need to be grounded psychologically. Here, we briefly consider what this means for the plausibility-as-proximity and plausibility-as-good-features claims.

*Plausibility-As-Proximity*². Many techniques argue that the instances in the CF need to be close [Wachter *et al.*, 2018]. But, aligning similarity with plausibility is problematic. First, there is no concrete evidence – as in human ratings studies --

showing that plausible CFs are more similar. Second, reporting that the counterfactuals generated involve highly-similar pairs invites the question: “how close do they have to be, to be plausible?”. A CF with a low distance-score could well be incomprehensible if it violates common-sense (e.g., a house with 2.312 rooms might be very close to the target house of 2 rooms but meaningless; see also Figure 2). Third, in computational evaluations a wide range of different distance metrics are used, which raises the question as to which one is the right one (is it L_1 , L_2 or some other variant)? So, though papers report distance metrics for computational evaluations because these metrics are not grounded psychologically, they do not tell us whether people will find them acceptable. We know of no user studies that confirm which distance metric is a good proxy for the psychological distance of CF explanations (e.g., something akin to Wang *et al.*’s, [2004] structural similarity index (SSIM) for images). If similarity is to be used as a proxy for plausibility, the chosen metric needs to be psychologically grounded (and explicitly, linked to people’s assessments of plausibility).

Plausibility As More-Good-Features. Others argue for the importance of “good” features for plausibility. But, aligning the plausibility of CFs with the appropriate use of “more good features” is also problematic. First, as with similarity, we have no direct evidence on the link between plausibility and

¹ For now, we will set aside the circularity of this sort of evaluation, but let’s note that it is an issue to be addressed.

² Note, the concept of *plausibility* is not well understood in Psychology; though, there is agreement that it tends to depend

heavily on user’s knowledge of a domain, rather than on similarity *per se* [Connell and Keane, 2007].

the feature-characteristics of CFs. Though, intuitively, we may agree that immutable features should not occur in plausible CFs, we do not know how different features-types – actionable, mutable, feasible or causal -- impact people’s perception of explanations; for example, how many “good” features are required for a plausible CF or how does the acceptability of explanations change with variations in a user’s causal model of the domain and so on. Note, that [Lakkaraju and Bastani, 2020] recently found that people can be *mised* into trusting a model more if the factual explanations used avoid “sensitive” features (such as gender and race). So, without thorough user studies, the plausibility of CFs with more-good-features remains an open question.

5 Deficit #3: The Shape of Sparsity

It is generally agreed that “good” CF explanations will be *sparse* (*aka* have few feature differences). This sparsity claim is made on intuition, sometimes with a hand-wave to psychological research on working memory limitations, visual memory, or limits on human category learning. An automated explanation for a house-price prediction saying “if it had 2 more rooms and a bigger garden, it would be double the price” is intuitively viewed as being better than one saying “if it had 2 more rooms, a bigger garden, a larger bathroom, a wider entrance-hall and a two-car garage, it would be double the price”. But, how sparse should sparse be? Most researchers do not commit to an optimal sparsity level for “good” CFs. The few papers that do commit, vary widely in their positions, from 1-2 feature-differences [Schleich *et al.*, 2021; Keane and Smyth, 2020] to 8-30 feature-differences [Martens and Provost, 2014]. Choosing a number seems to be important as it looks like many popular methods do not find low-sparsity CFs [Schleich *et al.*, 2021]. However, until recently, there was little hard evidence on what sparsity level might be psychologically optimal.

All the user studies that show some the general effect of CFs in XAI [Dodge *et al.*, 2019; Lim *et al.*, 2020] use single-feature-difference CFs, so they are silent on the impact of >1 feature-differences on people’s responses. The first direct tests of sparsity come from a single group in the University of Ulm [Förster *et al.*, 2020a, 2020b]. Over a series of studies, assessing CF methods, they systematically varied the median feature-differences in CF explanations shown to users. Interestingly, they found that people do not prefer 1-difference CFs but do find CF explanations with 2-3 feature-differences to be much better.

The point of this discussion is not all about what the optimal psychological sparsity-level might be, but is rather again to underscore the importance of using evaluation metrics that are backed by user studies. Clearly, the precise optimal sparsity level for people is likely to vary (e.g., in tabular versus text/image data, especially when chunking is taken into account)³; however, it is probably wise to settle on

a low-end score, if only to set a high bar for distinguishing between different methods.

6 Deficit #4: Covering Coverage

Any adequate counterfactual method needs to guarantee that, on the whole, it will produce “good” counterfactual explanations and avoid “bad” ones over some set of representative problems. If a method generates implausible explanations, even some of the time, given people’s tendency towards “algorithmic aversion” it will fail [Burton *et al.*, 2019]. Laugel *et al.* [2019] showed that one type of “bad” CF (i.e., out-of-distribution items) can be as high as 36% for some CF methods and Delaney *et al.* [2020] have shown that even close, low-sparsity CFs can be out-of-distribution (see Figure 2). In the 100 systems reviewed here, we found that only 22% report “coverage results”, though the definitions of the concept differ [Keane and Smyth, 2020; Schleich *et al.*, 2021; Dandl *et al.*, 2020].

What we are calling *coverage* can be measured in a number of different ways. Some have used out-of-distribution (OOD) measures to track the numbers of invalid CFs being produced by models: including, using IM1 and IM2 [Van Looveren and Klaise, 2019], Local Outlier Factor (LOF) [Breunig *et al.*, 2000] and One-Class-SVM (OCSVM) [Schölkopf *et al.*, 1999]. Obviously, generative CF methods place a strong emphasis on staying within the distribution and, as such, have championed OOD evaluations [Joshi *et al.*, 2019; Liu *et al.*, 2019; Singla *et al.*, 2019]. However, these measures tell us more about “bad” CF explanations than they do about “good” ones. If we assume OOD-CFs are implausible (and this needs to be verified psychologically) then the remaining in-distribution CFs only “may” be plausible; being in-distribution is not in itself a guarantee of plausibility (see Figure 2). We need to know which of these in-distribution CFs that *are* plausible.

Keane and Smyth [2020] have proposed the idea of *explanatory competence* (by analogy to *predictive competence* [Juarez *et al.*, 2018]) using a definition of *explanatory coverage* (*XP_Coverage*). Assume we have a function that captures psychologically-acceptable CF explanations -- *explains*(*c*, *c'*) -- where, *c*, is the test instance and *c'* the CF-instance. Then, the *explanatory competence* of a dataset, *C*, can be represented by a *coverage set* (Eq. 1) and degree of explanatory competence is the size of the coverage set as a fraction of the dataset (Eq. 2):

$$\underline{XP_Coverage_Set}(C) = \{c' \in C \mid \exists c \in C - \{c\} \ \& \ \underline{explains}(c, c')\} \quad (1)$$

$$\underline{XP_Coverage}(C) = \underline{XP_Coverage_Set}(C) / |C| \quad (2)$$

This evaluation metric is critical as it estimates the likelihood that users will encounter a good explanation when using a given method. However, an issue for this measure is how to

³ Note, what counts as a feature will not always be an input-feature (as in tabular data); we may be dealing with latent or

“semantic” features of the model, which may be *superpixels* for image data or *motifs* in time-series, and so on.

define the *explains* function. Keane and Smyth [2020] adopted the simple expedient of defining it as any CF with ≤ 2 feature-differences; but, more complex, psychologically-backed definitions need to be defined. We argue this measure needs to be part of any standard, benchmark evaluation.

7 Deficit #5: Comparative Testing

The final deficit to be noted about the CF literature is the lack of comparative testing in CF papers. We found only 40% (40 of 100 method papers) reported any form of comparative testing⁴. While the community has clearly moved beyond John McCarthy’s “Look Ma, no hands !” phase, given the sheer number of different methods now in the literature, more comparative testing is clearly needed. Fortunately, more papers recently tend to report such tests and many groups are making their code publicly available. We believe that with an agreed set of evaluation measures that are grounded in user tests, this field is poised to deliver solid scientific advances.

8 Roadmapping & Benchmarking

The human aspect of XAI research places a whole new set of constraints on the AI methods being developed to ensure fairness, accountability and trustworthiness. In this survey, we reviewed the evaluative shortcomings of XAI research on counterfactual explanations, mainly with an eye to providing a better psychological grounding for the area. To cement real progress in this area, we recommend a roadmap for future work and set of evaluative benchmarks to be adopted.

8.1 Roadmap for Psychological Grounding

We have argued that the core deficit facing CF XAI is the gap in user testing and the proper psychological grounding of computational evaluations. As such, our research roadmap recommends a program of general and specific testing of counterfactual explanations.

Broad user-testing needs to be carried across a range of diverse domains (e.g., decision making systems using datasets in different enterprise domains). For these general tests, we recommend the studies by [Dodge *et al.*, 2019] and [Lage *et al.*, 2020] as excellent experimental designs.

Specific user testing is also required to backstop the benchmark computational metrics to be used for comparative evaluations of methods. These studies will need to determine (i) the relationship between proximity and “good” CF explanations as determined by human-users and the most suitable distance metric to approximate this (whether it be L_1 , L_2 , or some other metric), (ii) the bounds on sparsity (to determine the optimal frequencies in feature-differences for comprehensibility, in different domains), (iii) how different classes of feature-differences are cognitively appraised (e.g., mutable/non-mutable, causal, actionable) and how this interacts with expertise in a domain, (iv) whether people can spot out-of-distribution CFs, how they appraise them and the

cognitive factors affecting these appraisals. For these specific tests, we recommend [Förster, *et al.*, 2020a, 2020b] and [van der Waa *et al.*, 2021] for the best experimental designs.

8.2 Benchmarking Evaluative Methods

As part of this roadmap we also need a standardized set of “proxy” computational evaluation metrics to help us decide between the 100-odd methods in the literature. We propose four benchmarking metrics that are selected to address the key insights guiding CF XAI research: including, benchmark metrics for proximity, sparsity, coverage and relative distance. We do not list benchmark datasets as there is already good agreement across papers on the key ones to use.

Proximity. Distance metrics are a rough, but reasonable, proxy for the overall performance of a CF method, though they can hide a lot; by convention, the literature has commonly used L_1 - and L_2 -norms. To enable retrospective comparisons the L_1 -norm probably has to be used, alongside the L_2 -norm, until it is clear which one is the more psychologically-valid measure. Karimi *et al.* [2020b] provide a good example of this sort of evaluation.

Sparsity. We have seen that sparsity is a critical index of the psychological acceptability of CF explanations. As distance metrics report averages, tests need to be broken out by sparsity levels showing the frequency of CFs generated; the range from 1-5 feature-differences seems critical, as it may well discriminate models (e.g., a model producing most of its CFs with >4 differences on tabular data may be suspect,). Schleich *et al.* [2021] give a good example of such reporting.

Coverage. Armed with a definition of “good counterfactuals” (even a rough one based on low-sparsity), a coverage metric capturing the proportion of good CFs for test-sets will provide a global estimate of the adequacy of methods; showing how likely they are to generate unacceptable CFs. Smyth and Keane [2021] provide a good example of this sort of reporting. Allied to this, we also need a measure of “bad” CFs being generated, using an OOD measure; we recommend the Local Outlier Factor (LOF) metric, for now, used in some papers [Breunig *et al.*, 2000].

Relative Distance. Finally, we propose a relative-distance measure (with the most psychologically-valid distance metric) comparing the mean distance of CF-pairs (between the test and CF instance) over the mean distance of “native counterfactuals” (NUNs). This metric is included to assess the instance-guided insight, as it shows whether the CFs generated by a method are closer than “natural” CFs in the dataset, on the assumption (to be tested) that this makes them better (see [Smyth and Keane, 2021] for an example).

8.3 Caveats & Conclusions

We should conclude with three caveats about these proposals that are important. First, we recognize that there are wider issues about the use and assessment of CF algorithms that are not considered here; namely, issues around the fidelity of

⁴ Note, this number overestimates comparative tests as it includes papers that solely test variants of their own algorithm.

explanations to a model's predictions, the robustness of methods, the speed/complexity of the algorithm, ethical and stakeholder issues and so on (see [Sokol and Flach, 2020], for a long list of requirements). Here, we have focused on, what we see as, core issues about the psychological and computational evaluation of these methods. All of these other requirements are also important.

Second, we are conscious that we have *not* proposed a benchmark metric for featural aspects of CF explanations; for example, whether the CFs generated by a method observe mutability, actionability and causal constraints. This aspect of CF methods also needs to be evaluated in a standardized way but there is less agreement in the literature on how this might be done. Furthermore, any broadly-applicable metric will require a dataset-by-dataset agreement on lists of immutable/mutable features, actionable features and also, possibly, causal models. While it is not impossible to do this, it does require greater community agreement to work. For potential candidates for such evaluations see [Ustun *et al.*, 2019; Karimi *et al.*, 2020b; Schleich *et al.*, 2021].

Third and finally, we do not underestimate how difficult it will be to rectify the deficits identified here and how much work will be required by inter-disciplinary teams to address these issues. However, if XAI does not address these sorts of issues then the emerging impediments to the widespread deployment of AI systems will, simply, not be overcome. And, perhaps, more fundamentally, we will not see sufficient scientific progress in these parts of the AI firmament.

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 to the *VistaMilk SFI Research Centre* under Grant Number 16/RC/3835.

References

- [Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking inside the black-box. *IEEE Access*, 6:52138–52160, 2018.
- [Akula *et al.*, 2020] Arjun Akula, Shuai Wang, and Song-Chun Zhu. Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *AAAI-20*, pages 2594-2601, 2020.
- [Artelt and Hammer, 2020] André Artelt and Barbara Hammer. Convex density constraints for computing plausible counterfactual explanations. In *ICANN-20*, pages 353-365. Springer, 2020.
- [Anjomshoae *et al.*, 2019] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots. In *AAMAS-19*, 1078-1088, 2019.
- [Barocas *et al.*, 2020] , Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *FAT*20*, pages 80-89, 2020.
- [Breunig *et al.*, 2000] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jorg Sander. LOF: Identifying density-based local outliers. *SIGMOD Rec.*, 29:93–104, 2000
- [Burton *et al.*, 2019] Jason W. Burton, Mari-Klara Stein and Tina Jensen. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33: 220-239, 2019
- [Byrne, 2019] Ruth Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI-19*, pages 6276–6282, 2019.
- [Chou *et al.*, 2021] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and Causability in Explainable Artificial Intelligence. *arXiv preprint arXiv:2103.04244* (2021).
- [Connell and Keane, 2006] Louise Connell and Mark T. Keane. A model of plausibility. *Cognitive Science*, 30: 95-120, 2006.
- [Dandl *et al.*, 2020] Susanne Dandl, Christoph Molnar, Martin Binder and Bernd Bischl. Multi-objective counterfactual explanations. In *PPSN 2020*, pages 448-469. Springer, 2020.
- [Delaney *et al.*, 2020] Eoin Delaney, Derek Greene and Mark T. Keane. Instance-Based Counterfactual Explanations for Time Series Classification. *arXiv preprint arXiv:2009.13211* (2020).
- [Dhurandhar *et al.*, 2018] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam and Payel Das. Explanations based on the missing. In *NIPS 2018*, pages 592-603. 2018.
- [Dodge *et al.*, 2019] Jonathan Dodge, Vera Liao, Yunfeng Zhang, Rachel Bellamy and Casey Dugan. Explaining models. In *IUI-19*, pages 275-285, 2019.
- [Förster *et al.*, 2020a] Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. Fostering Human Agency. In *ICIS-2020*, paper 1963, 2020.
- [Förster *et al.*, 2020b] Maximilian Förster, Philipp Hühn, Mathias Klier, and Kilian Kluge. Capturing Users' Reality. In *HICSS-20*, p. 1274, 2020
- [Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.
- [Gomez *et al.*, 2020] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. ViCE: visual counterfactual explanations for machine learning models. In *IUI-20*, pages 531-535, 2020.
- [Goyal *et al.*, 2019] Yash Goyal, Ziyang Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee.

- Counterfactual visual explanations. In *ICML-19*, pages 2376-2384, 2019.
- [Hendricks *et al.*, 2018] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell and Zeynep Akata. Generating Counterfactual Explanations with Natural Language. In *ICML Workshop on HIML*, pages 95-98, 2018.
- [Hoffman and Zhao, 2020] Guy Hoffman and Xuan Zhao. A Primer for Conducting Experiments in Human–Robot Interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10: 1-31, 2020.
- [Juarez *et al.*, 2018] Jose Juarez, Susan Craw, J. Ricardo Lopez-Delgado and Manuel Campos. Maintenance of case-bases: Current algorithms after fifty years. In *IJCAI-18*, pages 5457-5463, 2018.
- [Joshi *et al.*, 2019] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615* (2019).
- [Kanamori *et al.*, 2020] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi and Hiroki Arimura. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *IJCAI-20*, pages 2855-2862, 2020.
- [Karimi *et al.*, 2020a] Amir-Hossein Karimi, Gilles Barthe, B Schölkopf and I Valera. A survey of algorithmic recourse. *arXiv preprint:2010.04050*, 2020.
- [Karimi *et al.*, 2020b] Amir-Hossein Karimi, Gilles Barthe, Borja Balle and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *AISTAT-2020*, pages 895-905, 2020.
- [Karimi *et al.*, 2020c] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge. *NeurIPS 2020*, 33, 2020.
- [Keane and Smyth, 2020] Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them. In *ICCB-20*, pages 163-178. Springer, 2020.
- [Kenny and Keane, 2019] Eoin M. Kenny and Mark T. Keane. Twin-systems for explaining ANNs using CBR. In *IJCAI-19*, pages 2708-2715, 2019.
- [Kenny and Keane, 2021] Eoin Kenny and Mark Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *AAAI-21*, 2021.
- [Kenny *et al.*, 2021] Eoin M. Kenny, Courtney Ford, Molly Quinn and Mark T. Keane. Post hoc explanations for deep learning. *Artificial Intelligence*, 294, 2021.
- [Lage *et al.*, 2019] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *HCOMP-19*, 59-67, 2019.
- [Lakkaraju and Bastani, 2020] Himabindu Lakkaraju and Osbert Bastani. How do I fool you? In *AAAI/ACM AIES-2020*, pages 79-85, 2020.
- [Laugel *et al.*, 2019] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard and Marcin Detyniecki. The dangers of post-hoc interpretability. In *IJCAI-19*, pages 2801-2807, 2019.
- [Leavitt and Morcos, 2020] Matthew L. Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.
- [Lewis, 2013] David Lewis. *Counterfactuals*. Wiley, 2013.
- [Lim *et al.*, 2009] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *CHI-09*, pages 2119-2128, 2009.
- [Liu *et al.*, 2019] Liu, Shusen, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. *arXiv preprint arXiv:1907.03077*, 2019.
- [Lucic *et al.*, 2020] Ana Lucic, Hinda Haned and Maarten de Rijke. Contrastive Local Explanations for Retail Forecasting. In *FAT*20*, pages 90-98, 2020.
- [Lundberg and Lee, 2017] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.
- [Martens and Provost, 2014] David Martens and Foster Provost. Explaining data-driven document classifications. *Mis Quarterly*, 38:73-100, 2014.
- [Miller, 2019] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence*, 267 (2019): 1-38.
- [Mittelstadt *et al.*, 2019] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in AI. In *FAT-19*, pages 279–288, 2019.
- [Mothilal *et al.*, 2020] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT*2020*, pages 607– 617, 2020.
- [Mueller *et al.*, 2019] Shane T. Mueller, Robert R. Hoffman, W. J. Clancey, A. K. Emery and G. Klein. Explanation in Human-AI Systems, 2019.
- [Nugent *et al.*, 2009] Conor Nugent, Dónal Doyle, and Pádraig Cunningham. Gaining insight through case-based explanation. *Journal of Intelligent Information Systems*, 32:267-295, 2009.
- [Pedreschi *et al.*, 2019] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri and Franco Turini. Meaningful explanations of black box AI decision systems. In *AAAI-19*, pages 9780-9784, 2019.
- [Poyiadzi *et al.*, 2020] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijn De Bie and Peter Flach. FACE:

- feasible and actionable counterfactual explanations.
In *HCOMP-20*, pages 344-350, 2020.
- [Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. Why should I trust you?
In *SIGKDD-16*, pages 1135-1144, 2016.
- [Russell, 2019] Chris Russell. Efficient search for diverse coherent explanations. In *FAT-19*, pp. 20-28, 2019.
- [Schleich *et al.*, 2021] Maximilian Schleich, Zixuan Geng, Yihong Zhang, and Dan Suciu. (2021). GeCo. arXiv preprint *arXiv:2101.01292*.
- [Schölkopf *et al.*, 1999] Bernhard Schölkopf, Robert C. Williamson, Alexander J. Smola, John Shawe-Taylor and John C. Platt. Support vector method for novelty detection. *NIPS-99*, pages 582-588, 1999.
- [Singla *et al.*, 2019] Sumedha Singla, Brian Pollack, Junxiang Chen and Kayhan Batma. Explanation by progressive exaggeration." In *ICLR-19*, 2019.
- [Smyth and Keane, 2021] Barry Smyth and Mark Keane. A Few Good Counterfactuals. *arXiv:2101.09056*, 2021.
- [Sokol and Flach, 2020] Kacper Sokol and Peter Flach. Explainability fact sheets. *FAT*20*, pp. 56-67. 2020.
- [Ustun *et al.*, 2019] Berk Ustun, Alexander Spangher and Yang Liu. Actionable recourse in linear classification. In *FAT-19*, pages 10-19, 2019.
- [van der Waa *et al.*, 2021] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers and Mark Neerinx. Evaluating XAI. *Artificial Intelligence*, 291, 2021.
- [Van Looveren and Klaise, 2019] Van Looveren, Arnaud, and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.
- [Verma *et al.*, 2020] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A Review. *arXiv preprint arXiv:2010.10596*, 2020.
- [Wachter *et al.*, 2018] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2018.
- [Wang *et al.*, 2004] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13:600-612, 2004
- [Woodward, 2005] Woodward, James. *Making things happen: A theory of causal explanation*. Oxford University Press, 2005.