GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations

Dan Ley [*1] **Saumitra Mishra** [2] **Daniele Magazzeni** [2]

Abstract

Counterfactual explanations have been widely studied in explainability, with a range of application dependent methods prominent in fairness, recourse and model understanding. The major shortcoming associated with these methods, however, is their inability to provide explanations beyond the local or instance-level. While many works touch upon the notion of a global explanation, typically suggesting to aggregate masses of local explanations in the hope of ascertaining global properties, few provide frameworks that are both reliable and computationally tractable. Meanwhile, practitioners are requesting more efficient and interactive explainability tools. We take this opportunity to propose Global & Efficient Counterfactual Explanations (GLOBE-CE), a flexible framework that tackles the reliability and scalability issues associated with current state-of-the-art, particularly on higher dimensional datasets and in the presence of continuous features. Furthermore, we provide a unique mathematical analysis of categorical feature translations, utilising it in our method. Experimental evaluation with publicly available datasets and user studies demonstrate that GLOBE-CE performs significantly better than the current state-of-the-art across multiple metrics (e.g., speed, reliability).

1. Introduction

Counterfactual explanations (CEs) construct input perturbations that result in desired predictions from machine learning (ML) models (Verma et al., 2020; Karimi et al., 2020; Stepin et al., 2021). A key benefit of these explanations is their ability to offer recourse to affected individuals in certain settings (e.g., automated credit decisioning). Recent years have witnessed a surge of subsequent research, identifying desirable properties of CEs (Wachter et al., 2018; Barocas et al., 2020; Venkatasubramanian & Alfano, 2020), developing the methods to model those properties (Poyiadzi et al., 2020; Ustun et al., 2019; Mothilal et al., 2020; Pawelczyk et al., 2021), and understanding the weaknesses and vulnerabilities of the proposed methods (Dominguez-Olmedo et al., 2021; Slack et al., 2021; Upadhyay et al., 2021; Pawelczyk et al., 2022). Importantly, efforts thus far have largely centered around generating explanations for individual inputs.

Such analyses can vet local model behaviour, though it is seldom obvious that any of the resulting insights would generalise globally. A local CE for a specific sample cannot represent the bias of the entire model i.e. it may suggest that a model is not biased against a protected attribute (e.g., race, gender), despite net biases existing. A potential way to gain such insights is to aggregate local explanations (Lundberg et al., 2020; Pedreschi et al., 2019; Gao et al., 2021), but since the generation of CEs is generally computationally expensive, it is not evident that such an approach would scale well or lead to *reliable* conclusions about a model's behaviour. Be it during training or post-hoc evaluation, global understanding ought to underpin the development of ML models *prior* to deployment, and reliability and efficiency play important roles therein. We seek to address this in the context of global counterfactual explanations (GCEs).

We define a GCE to be *a global direction* along which a group of inputs may travel to alter their predictions. In prior work, GCEs simply took the same form as CEs, with the only difference being that a GCE should be applied to an entire group of inputs. Unfortunately, such a formulation, though instinctive, fails to overcome the trade-off between coverage and cost. Our relaxed objective, where each GCE represents just the translation direction, successfully overcomes this limitation. Figure 1 demonstrates how individual inputs are endowed with variable magnitude, or cost, for one particular global counterfactual direction. Additionally, for reference, we distinguish *counterfactuals* (the altered inputs) from *counterfactual explanations* (any representation of the change required, e.g., increase salary by $5k, or "If status is 3, Then change status to 1", etc.).

[*]Work done as an intern with J.P. Morgan AI Research [1]Harvard University [2]J.P. Morgan AI Research. Correspondence to: Dan Ley <dley@g.harvard.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Contributions Section 4 proceeds to introduce our framework Global & Efficient Counterfactual Explanations (GLOBE-CE). Though not strictly bound to recourse, our framework has the ability, as in Actionable Recourse Summaries(AReS) (Rawal & Lakkaraju, 2020), to seek answers to the big picture questions regarding a model's recourses, namely the potential disparities between affected subgroups, i.e. *Do race or gender biases exist in the recourses of a model? Can we reliably summarise these?*

Research thus far assumes GCEs to be fixed translations or rules. We relax this objective, and instead represent each GCE with a fixed translation direction $\underline{\delta}$, multiplied by an input-dependent, scalar variable k_j (Figure 1). This is distinct from finding a local CE for each input, as such an approach would return an infeasibly large range of directions to be interpreted, while also requiring significant computation. In contrast, using our approach, while each GCE has a fixed direction, magnitudes are allowed to vary input-wise. A set of n GCEs is sought to collectively cover the inputs, where n is small enough to be interpretable.

For the case $n = 1$, Figure 1 demonstrates how, if the GCE is a single, rigid translation, one must sacrifice either *coverage* (the percentage of inputs with altered predictions) or *average cost* (the average difficulty associated with executing the change for just the altered predictions). For instance, to achieve 100% coverage, all points must be pushed over the boundary, yielding an average cost equal to the cost of the furthest input. By allowing magnitudes to vary, we mitigate this trade-off significantly. The approach is natural, and the array of magnitudes returned can be easily analysed.

To determine the direction of each translation, GLOBE-CE deploys a general CE algorithm flexible to various desiderata. These include but are not limited to sparsity, diversity, actionability and model-specific CEs (Section 2.1). The novelty of our contributions lies in a) varying k_j input-wise and b) proving that arbitrary translations on one-hot encodings (categorical data) can be expressed using If/Then rules. To the best of our knowledge, this is the first work that addresses mathematically the direct addition of translation vectors to one-hot encodings, at least in the context of CEs.

Section 5 evaluates the efficacy of our GLOBE-CE framework along the three fundamental dimensions of *coverage*, *average cost*, and *speed* (the time spent computing GCEs). We argue that GCEs that fail to attain maximum coverage or minimum cost can be misleading, raising concerns around the safety of ML models vetted by such explanations (Section 3). GLOBE-CE outperforms competing methods in terms of coverage and cost, while having an order of magnitude faster runtime. User studies comprising ML practitioners additionally demonstrate the ability of the GLOBE-CE framework to more reliably detect recourse biases where previous methods fall short.

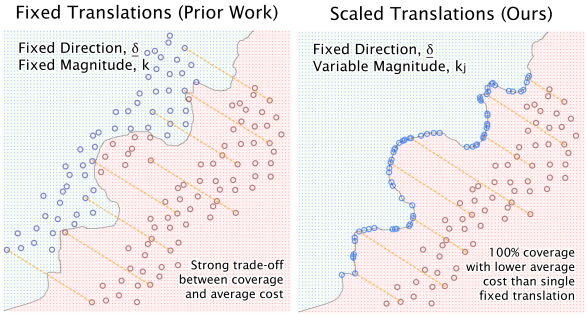

Figure 1. Blue and red regions indicate positive and negative model predictions. Red circles are inputs, and blue/red circles are inputs after translations (orange dotted lines) are applied. **Left:** Prior work assumes GCEs to be fixed translations/rules. The resulting trade-off between coverage and average cost was discussed in Kanamori et al. (2022). **Right:** We argue that a fixed direction, variable magnitude set-up can greatly improve performance while retaining interpretability. Each figure demonstrates how one type of GCE works to transfer points from the red region to the blue.

Concretely, our contributions are summarized as follows:

- We propose a framework that permits GCEs to have variable magnitudes while preserving a fixed translation direction. This relaxation mitigates the trade-off between coverage and cost that has been commonly accepted in existing research (Figure 1).
- We prove that arbitrary translations on one-hot encodings (categorical data) can be expressed using If/Then rules. To the best of our knowledge, this is the first work that addresses mathematically the direct addition of translation vectors to one-hot encodings.
- We demonstrate that GLOBE-CE outperforms competing methods in coverage, cost, and runtime (executing orders of magnitude faster across 4 benchmark datasets and 3 commonly used model types).

2. Related Work: Local and Global CEs

2.1. Local Counterfactual Explanations

Wachter et al. (2018) is one of the earliest introductions of CEs in the context of understanding black box ML models, defining CEs as points close to the query input (w.r.t. some distance metric) that result in a desired prediction. This inspired several follow-up works proposing desirable properties of CEs and presenting approaches to generate them. Mothilal et al. (2020) argues the importance of diversity, while other approaches aim to generate plausible CEs by considering proximity to the data manifold (Poyiadzi et al., 2020; Van Looveren & Klaise, 2021; Kanamori et al., 2020) or by accounting for causal relations among input

features (Mahajan et al., 2019). Actionability of recourse[1] is another important desideratum, suggesting certain features be excluded or limited (Ustun et al., 2019). In another direction, some works generate CEs for specific model categories, such as tree-based (Lucic et al., 2022; Tolomei et al., 2017; Parmentier & Vidal, 2021) or differentiable (Dhurandhar et al., 2018) models. Detailed surveys on CEs naturally follow (Karimi et al., 2020; Verma et al., 2020).

2.2. Global Counterfactual Explanations

Despite a growing desire from practitioners for global explanation methods that provide summaries of model behaviour (Lakkaraju et al., 2022), the struggles associated with summarising complex, high-dimensional models globally is yet to be comprehensively solved. For explanations in general, some manner of local explanation aggregation has been suggested (Lundberg et al., 2020; Pedreschi et al., 2019; Gao et al., 2021), though no compelling results have been shown that are both reliable and computationally tractable for GCEs specifically. Lakkaraju et al. (2022) also indicates a desire for increased interactivity with explanation tools, alongside reliable global summaries, but these desiderata cannot be paired until the efficiency issues associated with global methods are addressed in research.

Such works have been few and far between. Both Plumb et al. (2020) and Ley et al. (2022) have sought global translations which transform inputs within one group to another desired target group, though neither accommodate categorical features. Meanwhile, Becker et al. (2021) provides an original method, yet openly struggles with scalability. Gupta et al. (2019) attempts to equalize recourse across subgroups during model training, but with no framework for global interpretation. Rawal & Lakkaraju (2020) proposes AReS, a comprehensive GCE framework which builds on previous work (Lakkaraju et al., 2019). AReS adopts an interpretable structure, termed two level recourse sets. To our knowledge, only AReS and recent adaptation CET (Kanamori et al., 2022) pursue GCEs for recourse, yet the latter reports runtimes in excess of three hours for both methods.

3. Challenges: Reliability vs Efficiency

Since our goal is to provide practitioners with fast and reliable global summaries, potentially at each iteration of training (Gupta et al., 2019), we must take steps to bridge the gap between *reliability* (defined below) and *efficiency*. Even the most comprehensive works that target maximum reliability (Rawal & Lakkaraju, 2020; Kanamori et al., 2022) suffer computation times in excess of three hours on relatively small datasets such as German Credit (Dua & Graff, 2019). In parallel, there exists a body of research advocating strongly the use of inherently interpretable models for these cases, where performance is not compromised (Rudin, 2019; Rudin & Radin, 2019; Chen et al., 2018). If black-boxes are reserved for more complex scenarios, we should seek a GCE method that both executes efficiently and scales well, criteria by which current works have fallen short.

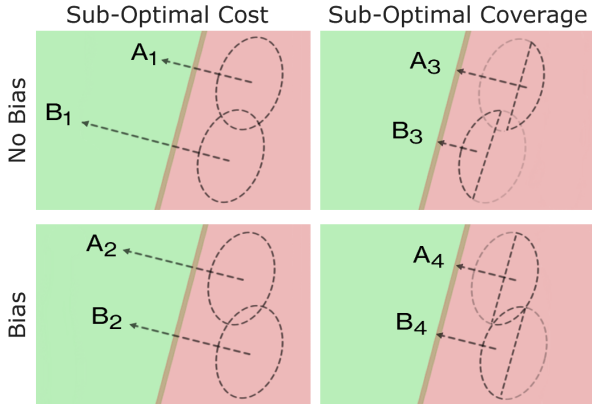

Figure 2. Common pitfalls with *unreliable* recourse bias assessment for GCEs A_i, B_i (ℓ_2 distance represents cost). Conceptual situations of bias/no bias vs sub-optimal cost/coverage displayed. Light dotted lines represent all inputs from either particular subgroup, A or B, while dark dotted lines are inputs within that subgroup for which the respective GCEs apply.

Reliability In this work, we define reliable GCEs to be those that can be used to draw accurate conclusions of a model's behaviour, and argue that this requires maximum coverage and minimum average costs. For instance, a model with higher recourse costs[2] for subgroup A than subgroup B is said to exhibit a *recourse bias* against subgroup A, and reliable GCEs would yield the *minimum* costs of recourse for either subgroup, such that this bias could be identified. If the costs found are sub-optimal, or GCEs only apply to a small number of points (sub-optimal coverage), the wrong conclusions might be drawn, as in GCEs A_i, B_i of Figure 2 e.g., in the top-left figure, both subgroups (shown as dotted circles) are equidistant from the decision boundary (no recourse bias), but since the recourse costs (length of A_1 and B_1) are different, one might incorrectly identify a bias.

Essentially, in the absence of minimum cost recourses, biases may be detected where not present (A_1, B_1) or not detected where present (A_2, B_2). Similarly, without sufficient coverage, the same phenomena may occur (A_3, B_3 and A_4, B_4, respectively). The further these metrics stray from optimal, the less likely any potential subgroup comparisons are of being reliable. We argue that maximising reliability

[1] Although their precise definitions differ, counterfactuals and recourse are considered the same in the context of this work.

[2] For CEs, cost is typically calculated as the ℓ_2 norm of the translation. For recourse, cost aims to calculate the difficulty associated with performing the translation.

thus amounts to maximising coverage while minimising recourse costs. The wider scope of reliability may encompass other desiderata as outlined in Section 2.1, though we refer their impact on bias assessment to future work.

Efficiency In our experiments, we define efficiency in relation to the average CPU time taken in computing GCEs.

4. Our Framework: Global & Efficient Counterfactual Explanations (GLOBE-CE)

We proceed to detail our proposed GLOBE-CE framework, where we discuss below: 1) the representation of GCEs that we choose, 2) theoretical results on categorical feature arithmetic, and 3) the GLOBE-CE algorithm.

4.1. Our Representation: Scaled Translation Vectors

We propose a novel and interpretable GCE representation: scaled translation vectors, as depicted in Figures 1 and 3. Simply put, for inputs that belong to a particular subgroup $\underline{x} \in \mathcal{X}$, we can apply a translation $\underline{\delta}$ with scalar k such that $\underline{x}_{\text{CF}} = \underline{x} + k\underline{\delta}$ is a valid counterfactual.[3] For each $\underline{x} \in \mathcal{X}$, our framework computes the respective minimum value of k required for recourse. The main appeal of this approach is its improvement with respect to the interpretability to performance trade-off that other methods suffer from. Using too few translations limits the performance of previous methods, yet large numbers of GCEs cannot easily be interpreted. Figure 1 demonstrates conceptually how one can achieve maximum coverage with a single translation at comparably lower average costs to previous methods which do not utilise variable magnitudes (Section 3 details why this is necessary for reliable bias assessment). The range of magnitudes for all inputs can be interpreted relatively easily (Section 5.1).

We posit that our method of a) assigning a single vector direction to an entire subgroup of inputs, b) travelling along this vector, and c) analysing the minimum costs required for successful recourses per input of the subgroup, is the natural global extension to one of the simplest forms of local CE: the *fixed* translation. In fact, the connection between local and global explanations may be more intimate than current research implies. Works suggesting to learn global summaries from local explanations (Pedreschi et al., 2019; Lundberg et al., 2020; Gao et al., 2021) and approaches suggesting to learn global summaries directly (Rawal & Lakkaraju, 2020; Kanamori et al., 2022; Plumb et al., 2020; Ley et al., 2022) tend to perceive global explanations as distinctly different challenges. Our framing of GCEs as a local problem is akin to treating groups of inputs as single instances, generating directions before subsequently, through scaling, efficiently capturing the variation within the set of local instances and their proximity to the decision boundary.

Unlike prior work, this implies that we can tackle a wide range of minimum costs, and potentially complex, non-linear decision boundaries, despite a fixed direction $\underline{\delta}$. AReS (Rawal & Lakkaraju, 2020) and recent similar work Counterfactual Explanation Trees (CET) (Kanamori et al., 2022) do not propose any such scaling techniques. The latter indicates the coverage to cost trade-off that occurs when one is limited by a *fixed* translation, yet eventually compromises for it. Other translation works (Plumb et al., 2020; Ley et al., 2022) do not utilise any form of scaling, and can be prone to failure since they target training data and not the model's decision boundary. In addition, neither provide steps to handle categorical features, an issue we address below.

4.2. Translations on Categorical Features

We assume one-hot encodings (else, these can be trivially encoded and decoded to suit) and provide, to the best of our knowledge, the first work in the context of CEs that reports mathematically the interpretation of a translation $k\underline{\delta}$ on a one-hot encoding (including the effects of scaling), deriving deterministic, interpretable rules from $k\underline{\delta}$. For examples of the interpretable rules sets generated by categorical translations, see Table 2 in Section 5.1, and Appendix B.2.

Theorem 4.1. *Regardless of the feature value of the input, any translation vector that is added to a one-hot categorical input can alternatively be expressed using If/Then rules, with just one unique Then condition.*

Proof (Sketch). Consider any one-hot encoded feature vector with feature labels ranging from 1 to n, denoted $\underline{f} = [f_1, f_2, ..., f_n] \in \{0,1\}^n$, where $|\underline{f}|_1 = 1$ and $F = \arg\max_i f_i$. Similarly, consider a translation vector of size n, denoted $\underline{\delta} = [\delta_1, \delta_2, ..., \delta_n] \in \mathcal{R}^n$, where $\Delta = \arg\max_i \delta_i$. The final vector post-translation is $\underline{g} = \underline{f} + \underline{\delta}$, and the final feature value is $G = \arg\max_i g_i$. Note $g_{i \neq F} = \delta_i$ and $g_F = \delta_F + 1$. We denote $g_G = \max_i g_i = \max(\delta_F + 1, \max_{i \neq F}(\delta_i))$. For $1 \leq F \leq n$, we now prove that if $G \neq F$ (i.e. a change in feature value occurs), we have the rule "If F, Then Δ".

In the case $F = \Delta$, $g_G = \max(\delta_\Delta + 1, \max_i(\delta_{i \neq \Delta})) = \delta_\Delta + 1$, as $\delta_\Delta = \max_i \delta_i$. Hence, $G = \Delta$ (no rule). In the case $F \neq \Delta$, $g_G = \max(\delta_F + 1, \delta_\Delta)$. Here, if $\delta_F + 1 > \delta_\Delta$, then $g_G = \delta_F + 1$ and $G = F$ (no rule). If $\delta_F + 1 < \delta_\Delta$, then $g_G = \delta_\Delta$ and $G = \Delta$, with the rule "If F, Then Δ".□

Theorem 4.2. *Regardless of the feature value of the input, any translation vector that is scaled by $k \geq 0$ and added to a one-hot categorical input can alternatively be expressed with the first m rules of a sequence.*

Proof (Sketch). Consider $\underline{f}$ and $\underline{\delta}$ defined in Theorem 4.1, and scalar k. For $i \neq \Delta$ and $k > 0$, Theorem 4.1 gives that

[3] Also denoted $round(\underline{x} + k\underline{\delta})$ in the presence of categorical features. The function $round(\underline{x} + k\underline{\delta})$ re-encodes one-hot outputs by selecting the largest feature values post-translation.

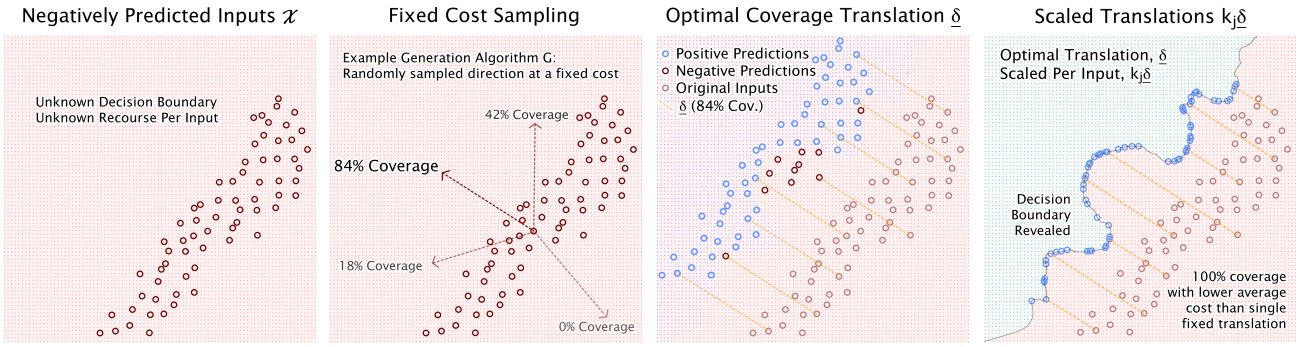

Figure 3. The GLOBE-CE framework (Algorithm 1) for an example generation algorithm G. Cost is ℓ_2 distance. **Left:** Negative predictions, $\mathcal{X}$. **Left Center:** We sample translations at a fixed cost, computing the coverage of each translation. **Right Center:** The translation with highest coverage is selected. **Right:** We scale $\underline{\delta}$ per input, returning the k_j value required for each input, where j indexes a vector of scalars $\underline{k}$. Theorems 4.1 and 4.2 bridge the gap between scaling translations and the discontinuous nature of categorical features.

$k\delta_i + 1 < k\delta_\Delta$ yields the rule "If i, Then Δ". Rearranging gives that if the lower bound $k > \frac{1}{\delta_\Delta - \delta_i}$ is satisfied, then the translation $k\underline{\delta}$ induces such a rule. Consider additionally the vector of lower bounds $\underline{k} = [k_1, k_2, ..., k_n] \in \mathcal{R}_+^n$ where $k_{i \neq \Delta} = \frac{1}{\delta_\Delta - \delta_i}$ and $k_\Delta = \infty$.

Lemma 4.2.1. *By inspection, we have that $k_i \leq k_m$ for any $i, m < n$ pair with $\delta_i \leq \delta_m$. As such, lower bounds for i and m are both satisfied if $k > k_m$. Thus, scaling $\underline{\delta}$ by $k > k_m$ induces both the rule corresponding to feature value m, and that of any other feature value i with $\delta_i \leq \delta_m$.*

For $k = 0$, we have no rules ($k\underline{\delta} = \underline{0}$). Let Δ_i now be the index of the ith smallest value in $\underline{\delta}$, such that $\Delta_1 = \arg\min_i \delta_i$ and $\Delta_n = \arg\max_i \delta_i = \Delta$. Thus, by Lemma 4.2.1, for $m < n$, we have that scaling $\underline{\delta}$ by $k_{\Delta_m} < k \leq k_{\Delta_{m+1}}$ induces rules for the first m feature values $\Delta_{1 \leq i \leq m}$. □

4.3. The GLOBE-CE Algorithm

Any particular CE may be represented with a *fixed* translation. The major contribution of the GLOBE-CE framework lies in the notion of *scaling the magnitudes* of translations. Though perhaps an uninteresting concept in the context of local CEs, when large numbers of inputs are present, scaling a translation $\underline{\delta}$ with an *input-dependent* variable k can efficiently solve for global summaries of the decision boundary. Figure 3 depicts the utility that a single, scaled translation vector can have. One can interpret a range of magnitudes, though cannot interpret a range of directions so easily; previous methods relied on using a small number of fixed GCEs.

Learning Translations In our setup, we learn explanations by adopting methods from instance-level CE research, generalising for any CE algorithm $G(B, \mathcal{X}, n)$ that considers, at a minimum, the model B being explained, the inputs requiring explanations $\mathcal{X}$, and the number n of returned GCEs $\underline{\delta}_1, \underline{\delta}_2, ..., \underline{\delta}_n = \Delta$. This enables previous CE research to simply be extended when seeking global explanations; we show in Section 5.2 how a simple method (constrained random sampling) can be adapted for our purposes. GLOBE-CE then scales the ith GCE $\underline{\delta}_i$ over a range of m scalars $\underline{k} = k_1, k_2, ..., k_m$, repeating over all $1 \leq i \leq n$ GCEs and returning the counterfactuals $\mathcal{X}'$, the predictions $\mathcal{Y}' \in \{0, 1\}^{n \times m \times |\mathcal{X}|}$ and costs $\mathcal{C} \in \mathcal{R}_{\geq 0}^{n \times m \times |\mathcal{X}|}$.

Algorithm 1 GLOBE-CE Framework

Input: B, $\mathcal{X}$, G, n, $\underline{k}$, *cost*

$\Delta = G(B, \mathcal{X}, n)$ ▷ Generate GCE Directions
for $1 \leq i \leq n$ **do** ▷ For all GCEs
  for $1 \leq j \leq |\underline{k}|$ **do** ▷ For all Scalars
    $\mathcal{X}'_{ij} = round(\mathcal{X} + k_j \underline{\delta}_i)$ ▷ Counterfactuals
    $\mathcal{Y}'_{ij} = B(\mathcal{X}'_{ij})$ ▷ Predictions
    $\mathcal{C}_{ij} = cost(\mathcal{X}, \mathcal{X}'_{ij})$ ▷ Costs
  end for
end for

Output: Counterfactuals $\mathcal{X}'$, Predictions $\mathcal{Y}'$, Costs $\mathcal{C}$ (For all Inputs $\mathcal{X}$, Translations Δ and Scalars $\underline{k}$)

In practice, one could cap n at the maximum number of GCEs end users can interpret. Algorithm 1 can terminate when coverage and cost plateau, or n is reached. Scalars $\underline{k}$ can be chosen with an upper limit on the cost of GCEs i.e. such that $\underline{k}$ simply ranges linearly from 0 to a point that yields this limit. This may vary per translation, a property not captured here, though one that is easily implementable. Given the speed of our method (Section 5), we deem $m = 1000$ to be appropriate. For categorical features, there exist specific scalar values one may use (Theorem 4.2).

5. Experiments

The possible representations, as well as the efficacy of the GLOBE-CE framework, are evaluated herein. Section 5.1

Table 1. Comparison of the AReS and GLOBE-CE algorithms, highlighting differences in methodology, feature handling, performance, and efficiency. The main differences include the handling of continuous features as well as the overall efficiency of both methods.

Comparison	AReS	GLOBE-CE
Algorithm	Generates hundreds/thousands of items $\mathcal{SD}$ Searches $\mathcal{SD}^3$ for valid triples, V Optimises V to select a smaller set of triples, R	Generates n GCE directions Scales each direction across all inputs Returns information on minimum cost per input
Continuous Features	Bins continuous features, displayed as If-Then rules (searches for combinations between commonly occurring bins)	Does not bin continuous features, displayed as addition/subtraction (no binning leads to performance improvements)
Categorical Features	Displayed as If-Then rules	We prove that (scaled) translations can also be expressed as If-Then rules
Performance	Lower coverage and higher average cost	Higher coverage and lower average cost
Efficiency	Computationally slow (hours for best performance)	Computationally fast (seconds)

demonstrates how the GLOBE-CE translations may be interpreted under various representations, while Section 5.2 comments on the the adaptability of our framework to existing CE desiderata, and the specifics of the algorithm G that we use to generate translations in our experiments.

Sections 5.3 provides an evaluation of the reliability and efficiency of our GLOBE-CE framework against the current state-of-the-art, and Section 5.4 details a user study involving 24 machine learning practitioners tasked with assessing recourse bias in models using either AReS (Rawal & Lakkaraju, 2020) or GLOBE-CE. The study also uncovers some important criteria for bias assessment.

Baselines We utilise AReS as a baseline. AReS is a rules based approach, while GLOBE-CE is a translation based approach. Both methods can compare recourses between subgroups to assess potential biases. AReS mines high-probability itemsets in the dataset, then combines these to produce triples (If-If-Then rules, e.g., If Sex = Male, If Salary $<$ \$50K, Then Salary $>$ \$50K). A set of thousands of triples is optimized to return a smaller number of interpretable triples. Differences are highlighted in Table 1.

We also introduce a separate implementation we call Fast AReS, aiming to mitigate some of the issues in the original method (notably its efficiency and its performance on continuous features). The details of the Fast AReS mechanisms are contained in Appendix C. Herein, it will be utilised and analysed solely as an additional baseline.

Datasets We employ four publicly available datasets to assess our methods: COMPAS (Larson et al., 2016), German Credit (Dua & Graff, 2019), Default Credit (Yeh & Lien, 2009) and HELOC (FICO, 2018). These predict recidivism, credit risk, payment defaults, and credit risk, respectively. AReS (Rawal & Lakkaraju, 2020) utilised COMPAS and German Credit; we include these in order to verify similar results. We also introduce the larger Default Credit and HELOC datasets, which include significant proportions of continuous features (full details in Appendix A.1).

Models We train 3 model types: Deep Neural Network (DNN), XGBoost (XGB), and Logistic Regression (LR). Best parameters for each dataset and model are chosen. We elect to train models on 80% of the data, unlike Rawal & Lakkaraju (2020). Given the class label imbalance in German Credit (30:70) and the size of the dataset ($N = 1000$), training AReS on 50% of the data is likely to yield only $0.5 \times 1000 \times 0.3 = 150$ negative predictions on which to construct recourses.

Computing Recourse Cost AReS suggests binning continuous features, and specifying the cost of moving between two adjacent bins to be 1. We bin continuous features into 10 equal intervals post-training. We also take the cost of moving from one categorical feature to another to be 1. To better model actionability (one's ability to enact the change suggested by a recourse) the relative costs of features requires further attention. In reality, estimating these is a complex task, and as such we do not focus on cost estimation in this paper. Further discussion is provided in Appendix E.

5.1. Interpreting Translation Directions

The manner in which explanations are portrayed depends on the nature of the data and/or the desire to compare recourses. We introduce straightforward interpretations of GCEs in continuous and categorical contexts.

Our scaling approach induces *coverage-cost profiles* as shown in Figure 4, Left, enabling selection of coverage/cost combinations and bias assessment. Intuitively, if the translation magnitude is 0 ($k = 0$), coverage and cost are also 0; as magnitude grows, more inputs cross the decision boundary, increasing coverage and average cost. Coverage/cost values can be chosen or compared, avoiding the the arduous process of choosing a hyperparameter that weights coverage/cost.

Table 2. Example *Cumulative Rules Chart (CRC)* for categorical features in the German Credit dataset, representing the optimal GLOBE-CE translation at 5 scalar values. Rules are cumulatively added (from top to bottom), resulting in an increase in coverage and cost.

Feature(s)	New Rule Added	New Inputs		All Inputs	
		Coverage	Cost	Coverage	Cost
Account Status	If F2, Then F4	+33.5%	1.00	33.5%	1.00
Account Status	If F3, Then F4	+2.5%	1.00	36.0%	1.00
Account Status	If F1, Then F4	+45.2%	1.00	81.2%	1.00
Telephone	If F2, Then F1	+2.5%	1.80	83.7%	1.02
Employment	If Not F4, Then F4	+10.2%	1.95	93.9%	1.12

We adopt standard statistical methods to convey *minimum costs*, corresponding to the minimum scalars required to alter each input's prediction (Figure 4, Right).

Continuous Data In our experiments, costs scale linearly as a particular translation is scaled, as in AReS, and we deem it interpretable to display solely *mean translations* alongside the illustration of minimum costs (Figure 4), an assertion supported by our user studies, where participants interpreted explanations for GLOBE-CE considerably faster than in AReS. Further details are provided in Appendix B.1.

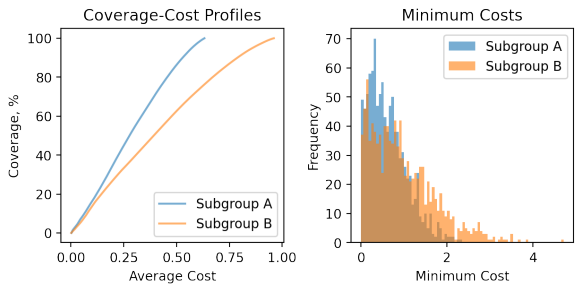

Figure 4. Example representation of costs, in the context of recourse bias assessment between two subgroups A and B (recourses are unfair on subgroup B). **Left:** Coverage-cost profiles (coverage increases with average cost). **Right:** Minimum costs per input.

Categorical Data Prior to our analysis, translations as raw vectors in input space lacked an immediate and intuitive interpretation on categorical data. Theorem 4.1 demonstrates that any translation can be interpreted as a series of If/Then rules, limited to one Then condition per feature, as portrayed by the individual rows in Table 2. Theorem 4.2 consequently proves that as a translation is scaled, If conditions are added to the rules for each feature (e.g., *Account Status* in Table 2). We name the resultant GCE representation a *Cumulative Rules Chart (CRC)*. See Appendix B.2 for further details.

5.2. Adapting the Generation Algorithm

To preface this section, we recognise that the scope of possible GCE generation algorithms is vast, and it should be stated that modifications to G along arbitrary criteria may impact efficiency in ways not investigated herein.

GCE Generation In the context of recourse, $G(B, \mathcal{X}, n)$ should deliver *diverse* and relatively *sparse* translations, targeting reliability (maximum coverage, minimum cost). We propose a specific generation algorithm $G(B, \mathcal{X}, n, n_s, c, n_f, p)$ for our experiments, which consists of uniform sampling of n_s translations at a *fixed* cost c. Additional parameters control the number of randomly chosen features n_f, and the power p to which random samples between 0 and 1 are raised, offering control over sparsity.

The n final GCEs are chosen to greedily maximise coverage; a new GCE direction has little performance utility unless it is significantly diverse from the previous directions. Thus, we take n as a proxy for diversity. In some cases, introducing more directions beyond $n = 1$ provides little performance improvement. There also exist model categories that provide alternative candidate G: model gradients in DNNs, feature attributions in XGB, and the mathematics of Support Vector Machines (SVMs) are explored in Appendix B.3.

5.3. GLOBE-CE Experimental Improvements

The GLOBE-CE explanations in our experiments are computed by the generation procedure posed in Section 5.2, and their interpretations are as detailed in Section 5.1. We test the performance (coverage, average cost and computation time) of both a single, scaled δ, as well as the diverse solution outlined in Section 5.2, with $n = 3$, denoting these GLOBE-CE and dGLOBE-CE, respectively. Translations are uniformally sampled at cost 2 and scaled in the range $0 \leq k \leq 5$, such that the maximum possible cost of a translation is 10 bins (the entire range for any one given feature), which we assume as a feasibility limit for recourse.

Specifying subgroups simplifies the complexity of the problem significantly, and so in our attempt to stress test these explanation methods and fully gauge their scalability, we will avoid this, and require the algorithms to compute GCEs for the entire input space in this section. Subgroup analysis is instead a core part of our user studies (Section 5.4).

Reliability Evaluation over 3 model types and 4 diverse datasets indicates that GLOBE-CE consistently exhibits the most reliable performance. Situations in which only moderate coverage is achieved by GLOBE-CE (e.g., COMPAS,

Table 3. Evaluating the reliability (coverage/cost) and efficiency of GLOBE-CE against AReS. Highlighted in red are GCEs that a) achieve below 10% coverage or b) require computation time in excess of 10,000 seconds (≈3 hours). Best metrics are shown in **bold**.

Models	Algorithms	Datasets											
		COMPAS			German Credit			Default Credit			HELOC		
		Cov.	*Cost*	*Time*	*Cov.*	*Cost*	*Time*	*Cov.*	*Cost*	*Time*	*Cov.*	*Cost*	*Time*
DNN	AReS	51%	2.31	101s	73%	1.6	2712s	7.22%	1.0	7984s	5.4%	1.0	9999s
	Fast AReS	64%	**1.45**	32.0s	72%	1.43	12.8s	99.8%	4.2	37.3s	52%	5.5	109.1s
	GLOBE-CE	66%	1.53	**7.08s**	85%	1.2	**2.28s**	98.5%	1.3	**3.6s**	93%	4.3	**4.66s**
	dGLOBE-CE	**70%**	1.46	9.15s	**90%**	**1.1**	2.63s	**100%**	**1.1**	7.86s	**95%**	3.8	5.46s
XGB	AReS	45%	1.9	205s	61%	1.5	2092s	11%	1.0	9999s	1.7%	1.0	9999s
	Fast AReS	83%	1.9	47.6s	65%	1.75	34.33s	93%	2.3	29.97s	28%	**2.1**	93.58s
	GLOBE-CE	78%	1.8	**9.61s**	**95%**	**1.02**	**5.04s**	96%	1.1	**2.94s**	58%	2.4	**4.7s**
	dGLOBE-CE	**91%**	**1.4**	12.4s	83%	1.03	5.95s	**100%**	**0.7**	6.35s	**80%**	2.4	5.6s
LR	AReS	79%	1.5	506s	85%	1.3	3566s	31%	1.2	9999s	4.8%	1.0	9999s
	Fast AReS	82%	1.7	43.0s	85%	1.3	9.3s	99%	2.1	17.82s	92%	1.6	127.3s
	GLOBE-CE	83%	1.20	**8.43s**	82%	**1.2**	**3.39s**	**100%**	**1.0**	**3.42s**	**100%**	**0.5**	**3.11s**
	dGLOBE-CE	**84%**	**1.18**	11.7s	**91%**	1.3	3.87s	**100%**	**1.0**	7.21s	**100%**	**0.5**	3.85s

DNN), we attribute to an inability of the model to provide recourses within our feasibility limit (after having exhaustively trialled solutions). Fast AReS achieves superior cost in one case (COMPAS, DNN), though with a significant drop in coverage (a natural trade-off).

Efficiency Applying AReS for several hours can yield coverage improvements, though we consider such time scales inappropriate given that a) we test relatively simple datasets (Section 3) and b) practitioners are requesting higher levels of interactivity in explainability tools (Lakkaraju et al., 2022). We argue too that the concept of sampling random translations at a fixed cost is more intuitive than tuning hyperparameters of terms associated with cost, and that GLOBE-CE thus grants a higher degree of interactivity, given its performance with respect to computation time.

5.4. User Studies

Extending our framework to provide comparisons between subgroups of interest, similarly to AReS, is a trivial matter of separating the inputs and learning translations separately.[4]

We conduct an online user study to analyse and compare the efficacy of GLOBE-CE and AReS in detecting recourse biases. The user study involves 24 participants, all with a background in AI and ML and some knowledge of post-hoc explainability. We include a short tutorial on CEs, GCEs, and the ideas of recourse cost and recourse bias. The study utilises two "black box" models: the first is a decision tree with a *model* bias against females, though with a *recourse* bias exhibited against males due to the nature of the data distribution (Figure 5); the second is an SVM (where theoretical minimum ℓ_2 costs can be computed as a ground truth) with a recourse bias against a *ForeignWorker* subgroup.

[4]However, it is recommended to generate the same set of n_s random translations for both subgroups as this eliminates random bias and executes faster. Alternatively, translations selected for each subgroup can be exchanged to directly compare recourses.

If Sex = Male:
- **If** Job = No **and** Property = No,
  Then Job = Yes **and** Property = Yes
- **If** Healthcare = No,
  Then Healthcare = Yes

If Sex = Female:
- **If** Job = No **and** Property = No **and** Savings = No,
  Then Job = Yes **and** Property = Yes **and** Savings = Yes
- **If** Healthcare = No,
  Then Healthcare = Yes

Figure 5. Depiction of Black Box 1, with *model* bias against females, yet *recourse* bias against males. 90% of rejected females satisfy the first rule with cost 3, and require healthcare with cost 1. In contrast, 90% of rejected males have healthcare, but require the first rule with cost 2, resulting in higher average recourse costs.

We randomly group the 24 participants into two equal subgroups, whereby each participant is presented with two global explanations generated from either AReS or GLOBE-CE. For each explanation, the user study asks two questions: 1) **do you think there exists bias in the presented recourse rules?**, and 2) **explain the reasoning behind your choice.** The first question is multiple choice with three answers, and the second is descriptive. Further, each study includes a 'filter' question to check if the participants understand the required concepts. We found that 3 participants in the first user study (for AReS) and 4 participants in the second user study (for GLOBE-CE) answered the filter question incorrectly. Hence, after removing the responses from these participants and ensuring we have equal participants in both studies, the number of participants whose responses we use for further analysis reduces to 16. Table 4 details the response breakdown. Further details of the models, recourses, and snapshots of the user study can be found in Appendix D.4.

Takeaway 1 The Black Box 1 study demonstrates that correct bias identification requires an understanding of the underlying distribution of recourses (the proportion of inputs each rule applies to, as described in Figure 5). GLOBE-CE provides distribution information (as in the Cumulative Rules Chart) of the recourses for the model in Figure 5. While AReS uncovers the biased *model*, it misleads users with regards to the biased *recourses* (7/8 participants claimed a recourse bias, though none described it correctly).

Takeaway 2 The Black Box 2 study demonstrates that sub-optimal cost yields misleading conclusions of recourse (in the way described in Section 3). GLOBE-CE outperforms AReS on cost with minimum recourse costs close to the ground truth SVM, and participants were able to more successfully uncover bias. Importantly, 7 of the 8 AReS users also found describing the recourse bias to be hard.

Table 4. Bias detection results from user studies. Bias and correct columns: number of users that identified a bias and number of users that described it correctly, respectively.

User Studies Breakdown	AReS		GLOBE-CE	
	Bias?	*Correct?*	*Bias?*	*Correct?*
Black Box 1	7/8	0/8	7/8	7/8
Black Box 2	1/8	0/8	5/8	4/8

6. Limitations & Future Work

We acknowledge the complexity of determining the costs associated with actioning specific counterfactuals. Not only could the costs of certain actions vary with time, depend on each other and themselves, or be susceptible to unknown degrees of randomness, but they could also depend on a multitude of factors surrounding the specific individual that is ultimately tasked with the action. Disregarding such difficulties, building a model that perfectly reflects a user's struggle in executing a particular set of actions might easily require an improper breach of said user's privacy. While Rawal & Lakkaraju (2020) proposes the Bradley-Terry model for fixed feature costs, it doesn't account for end-user properties or cost dependencies. We recognize that finding better cost models is a valuable research area, but due to the unsolved nature of actionability, we don't overly focus on cost estimation in this paper. Instead, we use unit costs between categorical features or per decile of continuous features (ℓ_1).

Second to this, a potential limitation is that our categorical translations always yield rules of the form "If A (or B or C...), Then X" for any one particular feature (with possible negation on the "If" term). However, it may also be useful to represent other forms, such as "If A, Then Not A" or "If A then (B or C)". The minimum costs yielded would not decrease as a result of these new forms, since our method can still represent them (e.g. "If A, Then B" would solve simply the first suggestion), though may have interpretability implications which could be explored in future work.

Finally, Appendix E provides an at length discussion of other remaining limitations and future work, including the relation of our method to current issues facing local CEs such as robustness (Dominguez-Olmedo et al., 2021) and sensitivity (Slack et al., 2021). The potential for GLOBE-CE to offer robust recourses can be explored in future work. Compared to previous methods, GLOBE-CE is likely to facilitate more efficient generation of robust counterfactuals, given its ability to easily adjust the distance into the decision boundary. This adaptability stems from its inherent capacity to scale counterfactual directions up or down. For instance, instead of scaling each GCE to reach the decision boundary, it could be scaled until some robustness criterion is met.

7. Conclusion

This work studies the current state of GCE methods, and addresses the issues associated with prior work. Fundamentally, these are a) requiring GCEs to be fixed-magnitude translations and b) computational complexity. With mounting desire from a practitioner viewpoint for access to fast, interactive explainability tools (Lakkaraju et al., 2022), it is crucial that such methods are not inefficient. We propose a novel GCE framework, GLOBE-CE, that further improves on the issues faced by the current state-of-the-art, AReS. Experiments with four public datasets and user studies demonstrate the efficacy of our proposed framework in generating accurate global explanations that assist in identifying recourse biases. In future we hope to apply the proposed GLOBE-CE approach to other real world use cases, and conceive of further generation algorithms that improve its performance beyond that demonstrated here. We hope that the work here inspires further research into the particularly under-studied area of GCEs, and proves useful as the development of explainability tools continues to grow.

Disclaimer This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Agrawal, R. and Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the International Conference on Very Large Data Bases*, VLDB '94, pp. 487–499, San Francisco, USA, September 1994. Morgan Kaufmann Publishers Inc. ISBN 1558601538. URL https://dl.acm.org/doi/10.5555/645920.672836.

Barocas, S., Selbst, A. D., and Raghavan, M. The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, January 2020. doi: 10.1145/3351095.3372830. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3503019. arXiv: 1912.04930.

Becker, M., Burkart, N., Birnstill, P., and Beyerer, J. A Step Towards Global Counterfactual Explanations: Approximating the Feature Space Through Hierarchical Division and Graph Search. In *Advances in Artificial Intelligence and Machine Learning*, 2021. URL https://publikationen.bibliothek.kit.edu/1000139219.

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., and Wang, T. An Interpretable Model with Globally Consistent Explanations for Credit Risk. In *NeurIPS Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*, Montréal, Canada, December 2018. URL https://arxiv.org/abs/1811.12615.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Advances in Neural Information Processing Systems*, Montréal, Canada, December 2018. URL https://proceedings.neurips.cc/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf.

Dominguez-Olmedo, R., Karimi, A.-H., and Schölkopf, B. On the Adversarial Robustness of Causal Algorithmic Recourse. *arXiv:2112.11313 [cs]*, December 2021. URL http://arxiv.org/abs/2112.11313. arXiv: 2112.11313.

Dua, D. and Graff, C. UCI Machine Learning Repository, 2019. URL http://archive.ics.uci.edu/ml.

Dutta, S., Long, J., Mishra, S., Tilli, C., and Magazzeni, D. Robust counterfactual explanations for tree-based ensembles. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5742–5756, Baltimore, USA, 17–23 Jul 2022. PMLR. URL https://proceedings.mlr.press/v162/dutta22a/dutta22a.pdf.

FICO. Explainable Machine Learning Challenge, 2018. URL https://community.fico.com/s/explainable-machine-learning-challenge.

Gao, J., Wang, X., Wang, Y., Yan, Y., and Xie, X. Learning Groupwise Explanations for Black-Box Models. In Zhou, Z.-H. (ed.), *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2396–2402, Virtual-Only, August 2021. International Joint Conferences on Artificial Intelligence Organization. URL https://doi.org/10.24963/ijcai.2021/330.

Gupta, V., Nokhiz, P., Roy, C. D., and Venkatasubramanian, S. Equalizing Recourse across Groups, 2019. URL https://arxiv.org/abs/1909.03166.

Kanamori, K., Takagi, T., Kobayashi, K., and Arimura, H. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2855–2862, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/395. URL https://www.ijcai.org/proceedings/2020/395.

Kanamori, K., Takagi, T., Kobayashi, K., and Ike, Y. Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1846–1870, Virtual-Only, March 2022. PMLR. URL https://proceedings.mlr.press/v151/kanamori22a.html.

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects. In *NeurIPS Workshop on ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)*, Virtual-Only, December 2020. URL https://ml-retrospectives.github.io/neurips2020/camera_ready/8.pdf. arXiv: 2010.04050.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. Faithful and Customizable Explanations of Black

Box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 131–138, Honolulu, USA, January 2019. Association for Computing Machinery. ISBN 9781450363242. URL https://doi.org/10.1145/3306618.3314229.

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., and Singh, S. Rethinking Explainability as a Dialogue: A Practitioner's Perspective, 2022. URL https://arxiv.org/abs/2202.01875.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lee, J., Mirrokni, V., Nagarjan, V., and Sviridenko, M. Non-Monotone Submodular Maximization under Matroid and Knapsack Constraints. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, Maryland, USA, May 2009. URL https://dl.acm.org/doi/10.1145/1536414.1536459. arXiv: 0902.0353.

Ley, D., Bhatt, U., and Weller, A. Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates. In *AAAI Conference on Artificial Intelligence*, Virtual-Only, February 2022. URL http://arxiv.org/abs/2112.02646.

Lucic, A., Oosterhuis, H., Haned, H., and de Rijke, M. FOCUS: Flexible Optimizable Counterfactual Explanations for Tree Ensembles. In *AAAI Conference on Artificial Intelligence*, Virtual-Only, February 2022. URL http://arxiv.org/abs/1911.12199. arXiv: 1911.12199.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. Explainable AI for Trees: From Local Explanations to Global Understanding. *Nature Machine Intelligence*, January 2020. URL https://www.nature.com/articles/s42256-019-0138-9.

Mahajan, D., Tan, C., and Sharma, A. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. In *NeurIPS Workshop on Do the Right Thing: Machine Learning and Causal Inference for Improved Decision Making*, Vancouver, Canada, December 2019. URL http://arxiv.org/abs/1912.03277. arXiv: 1912.03277.

Mishra, S., Dutta, S., Long, J., and Magazzeni, D. A survey on the robustness of feature importance and counterfactual explanations, 2021. URL https://arxiv.org/abs/2111.00358.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 607–617, Barcelona, Spain, January 2020. doi: 10.1145/3351095.3372850. URL https://dl.acm.org/doi/abs/10.1145/3351095.3372850. arXiv: 1905.07697.

Parmentier, A. and Vidal, T. Optimal Counterfactual Explanations in Tree Ensembles. In *International Conference on Machine Learning*, pp. 8422–8431, Virtual-Only, July 2021. URL http://proceedings.mlr.press/v139/parmentier21a/parmentier21a.pdf. arXiv: 2106.06631.

Pawelczyk, M., Bielawski, S., Heuvel, J. v. d., Richter, T., and Kasneci, G. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. In *Advances in Neural Information Processing Systems (Datasets & Benchmarks Track)*, August 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/b53b3a3d6ab90ce0268229151c9bde11-Paper-round1.pdf. arXiv: 2108.00783.

Pawelczyk, M., Datta, T., van-den Heuvel, J., Kasneci, G., and Lakkaraju, H. Algorithmic Recourse in the Face of Noisy Human Responses. *arXiv:2203.06768 [cs]*, March 2022. URL http://arxiv.org/abs/2203.06768. arXiv: 2203.06768.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F. Meaningful Explanations of Black Box AI Decision Systems. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 9780–9784, Honolulu, USA, January 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/5050.

Plumb, G., Terhorst, J., Sankararaman, S., and Talwalkar, A. Explaining Groups of Points in Low-Dimensional Representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7762–7771, Virtual-Only, July 2020. PMLR. URL http://proceedings.mlr.press/v119/plumb20a/plumb20a.pdf.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, New York, USA, February 2020. doi: 10.1145/3375627.3375850. URL https://dl.acm.org/doi/10.1145/3375627.3375850. arXiv: 1909.09369.

Rawal, K. and Lakkaraju, H. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12187–12198, Virtual-Only, December 2020. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/file/8ee7730e97c67473a424ccfeff49ab20-Paper.pdf.

Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, May 2019. URL https://www.nature.com/articles/s42256-019-0048-x.

Rudin, C. and Radin, J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, November 2019. doi: 10.1162/99608f92.5a8a3a3d. URL https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

Slack, D., Hilgard, A., Lakkaraju, H., and Singh, S. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*, volume 34, Virtual-Only, December 2021. URL https://proceedings.neurips.cc/paper/2021/file/009c434cab57de48a31f6b669e7ba266-Paper.pdf.

Stepin, I., Alonso, J. M., Catala, A., and Pereira-Fariña, M. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. In *IEEE Access*, volume 9, pp. 11974–12001, 2021. doi: 10.1109/ACCESS.2021.3051315. URL https://ieeexplore.ieee.org/document/9321372.

Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 465–474, Novia Scotia, Canada, August 2017. URL https://dl.acm.org/doi/10.1145/3097983.3098039.

Upadhyay, S., Joshi, S., and Lakkaraju, H. Towards Robust and Reliable Algorithmic Recourse, July 2021. URL http://arxiv.org/abs/2102.13620. arXiv: 2102.13620.

Ustun, B., Spangher, A., and Liu, Y. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19, Atlanta, USA, January 2019. doi: 10.1145/3287560.3287566. URL https://dl.acm.org/doi/10.1145/3287560.3287566. arXiv: 1809.06514.

Van Looveren, A. and Klaise, J. Interpretable Counterfactual Explanations Guided by Prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 650–665, Bilbao, Spain, September 2021. Springer. URL https://2021.ecmlpkdd.org/wp-content/uploads/2021/07/sub_352.pdf.

Venkatasubramanian, S. and Alfano, M. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 284–293, Barcelona, Spain, January 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372876. URL https://dl.acm.org/doi/10.1145/3351095.3372876.

Verma, S., Dickerson, J., and Hines, K. Counterfactual Explanations for Machine Learning: A Review, October 2020. URL http://arxiv.org/abs/2010.10596.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3063289. URL https://www.ssrn.com/abstract=3063289. arXiv: 1711.00399.

Yeh, I.-C. and Lien, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473–2480, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2007.12.020. URL https://www.sciencedirect.com/science/article/pii/S0957417407006719.

Appendix

This appendix is formatted as follows.

1. We discuss the *Datasets & Models* used in our work in Appendix A.
2. We discuss *Implementation Details* and *Example Outputs* of GLOBE-CE in Appendix B.
3. We discuss the *Implementation Details* for AReS and Fast AReS in Appendix C.
4. We list the *Experimental Details* of our work and analyse *Further Results* in Appendix D.
5. We discuss *Limitations* and areas of *Future Work* for GLOBE-CE in Appendix E.

Where necessary, we provide discussion for potential limitations of our work and future avenues for exploration.

A. Datasets & Models

Four benchmarked datasets are employed in our experiments, all of which describe binary classification and are publicly available. Details are provided below and in Table 5. Our experiments utilise three types of models: Deep Neural Networks (DNNs); XGBoost (XGB); and Logistic Regression (LR), described below and in Tables 6 through 8. Our user study also investigates linear kernel Support Vector Machines (SVMs), as these provide mathematical forms for minimum recourse.

A.1. Datasets

Where necessary, we augment input dimensions with one-hot encodings over necessary variables (e.g. Sex).

The **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) dataset (Larson et al., 2016) classifies **recidivism risk**, based off of various factors including race, and can be obtained from and is described at: `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`. This dataset tests performance of our method in low-dimensional and highly-categorical settings.

The **German Credit** dataset (Dua & Graff, 2019) classifies **credit risk** and can be obtained from and is described at: `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`. The documentation for this dataset also details a cost matrix, where false positive predictions induce a higher cost than false negative predictions, but we ignore this in model training. Note that this dataset, which tests mainly categorical settings, is distinct from the common Default Credit dataset, described hereafter.

The **Default Credit** dataset (Yeh & Lien, 2009) classifies **default risk** on customer payments and is obtained from and described at: `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`. This dataset stress-tests scalability (increased inputs, dimensions and continuous features).

The **HELOC** (Home Equity Line of Credit) dataset (FICO, 2018) classifies **credit risk** and can be obtained (upon request) from: `https://community.fico.com/s/explainable-machine-learning-challenge`. We drop duplicate inputs, and inputs where all feature values are missing (negative values), and replace remaining missing values in the dataset with median values. Notably, the majority of features are monotonically increasing/decreasing.

Table 5. Summary of the datasets used in our experiments. Although German Credit includes continuous features, we find that they have limited effect on the model both during training and in the resulting explanations.

Name	No. Inputs	Input Dim.	Categorical	Continuous	No. Train	No. Test
COMPAS	6172	15*	4	2	4937	1235
German Credit	1000	71*	17	3	800	200
Default Credit	30000	91*	9	14	24000	6000
HELOC	9871	23	0	23	7896*	1975*

*Denotes values post-processing (one-hot encoding inputs, dropping inputs).

A.2. Models

We train models with an 80:20 split for each dataset. While each model's parameters differ, the universal aims are to a) achieve sufficient coverage, b) avoid overfitting, and c) predict a similar class balance to the original data. The true proportion of negative labels in the training data of each dataset are **45%**, **30%**, **22%**, and **53%**, respectively. The Table 5 No. Train column dictates $|\mathcal{X}|$ in Tables 6 through 8, and the inputs in $|\mathcal{X}|$ with negative predictions from a particular model are denoted $|\mathcal{X}_{\text{aff}}|$.

Deep Neural Networks (DNNs) We use the common `PyTorch` library to construct and train DNNs. Widths and depths of such models are outlined in Table 6. Layers include dropout, bias and ReLU activation functions. We map the final layer to the output using softmax, and use Adam to optimise a cross-entropy loss function. Table 6 details various model parameters/behaviours.

Table 6. Summary of the DNNs used in our experiments.

Name	Width	Depth	Dropout	Train Cov.	Test Acc	$\lvert\mathcal{X}_{\text{aff}}\rvert$	$\lvert\mathcal{X}_{\text{aff}}\rvert/\lvert\mathcal{X}\rvert$
COMPAS	30	5	0.4	68%	65%	2552	**52%**
German Credit	50	10	0.3	84%	78%	243	**30%**
Default Credit	80	5	0.3	81%	81%	5232	**22%**
HELOC	50	5	0.5	74%	74%	4334	**55%**

XGBoost (XGB) Implementation from the common `xgboost` library.

Table 7. Summary of the XGB models used in our experiments.

Name	Depth	Estimators	γ, α, λ	Train Cov.	Test Acc	$\lvert\mathcal{X}_{\text{aff}}\rvert$	$\lvert\mathcal{X}_{\text{aff}}\rvert/\lvert\mathcal{X}\rvert$
COMPAS	4	100	1, 0, 1	70%	68%	2008	**41%**
German Credit	6	500	0, 0, 1	95%	74%	197	**25%**
Default Credit	10	200	2, 4, 1	90%	83%	3744	**16%**
HELOC	6	100	4, 4, 1	77%	74%	4323	**55%**

Logistic Regression (LR) Implementation from the common `sklearn` library.

Table 8. Summary of the LR models used in our experiments. Class Weights refer to the loss function used.

Name	Max Iterations	Class Weights (0:1)	Train Cov.	Test Acc	$\lvert\mathcal{X}_{\text{aff}}\rvert$	$\lvert\mathcal{X}_{\text{aff}}\rvert/\lvert\mathcal{X}\rvert$
COMPAS	1000	1:1	67%	65%	1940	**39%**
German Credit	1000	1:1	79%	76%	180	**23%**
Default Credit	2000	0.65:0.35	81%	83%	3858	**16%**
HELOC	2000	1:1	73%	75%	4282	**54%**

B. Our Framework: Global & Efficient CEs

This Appendix discusses several specificities concerning our methodology. Similarly to our AReS implementations, we acknowledge that there does exist scope to improve upon the efficiency of our method, though are encouraged by the superior results GLOBE-CE achieves relative to baselines.

B.1. Example GLOBE-CE Representations for Continuous Data (HELOC)

This section expands upon the various outputs of our framework (Section 5.1) in the context of continuous data from the HELOC dataset (FICO, 2018). We consider the following useful representations of the information computed in Algorithm 1: coverage-cost profiles, minimum costs, and mean translations.

Coverage-Cost Profiles The natural trade-off associated with CEs is that between coverage and cost; lower cost CEs are likely to cover fewer inputs, and vice versa. The literature typically accounts for this by introducing a hyperparameter (e.g. λ) to tune the trade-off. This however assumes a linear relationship between the metrics i.e. λ is held constant during optimisation. Not only is this unrealistic, as the relative importance of coverage normally varies with cost, but it introduces tuning difficulties for practitioners, i.e., it must be determined either *a priori* or through hyperparameter search which coverage-cost combination is optimal, and subsequently which λ value maps to this particular combination (non-trivial). As stated in Section 5.1, our scaling approach instead induces *coverage-cost profiles* as in Figure 1, providing a far more interpretable method for coverage/cost selection.

Minimum Costs & Mean Translations Standard statistical methods can be adopted to convey the *minimum costs* per input (corresponding to the minimum scalar required to alter the prediction). For continuous data, knowing that costs scale linearly as a particular translation is scaled, we deem it interpretable to display solely *mean translations* alongside the illustration of minimum costs, an assertion supported by our user studies, where participants interpreted GLOBE-CE explanations considerably faster than in AReS. These explanation representations are depicted in Figure 1.

B.2. Example GLOBE-CE Representations for Categorical Data (German)

We now proceed to examine typical GLOBE-CE outputs for categorical features, using the German Credit (Dua & Graff, 2019) dataset. Of particular interest is the representation of scaled categorical translations. The coverage-cost profile and minimum cost histogram naturally follow.

Cumulative Rules Charts for Scaled Translations For the sake of our analysis, consider purely the categorical features from German Credit (see Section 5.1 for the introduction to CRCs).

Recall that, under a given translation direction, there exists a minimum cost, or equivalently a minimum scalar, at which recourse is achieved (shown in blue in Figure 6), if indeed recourse can be achieved at all. However, unlike continuous features, representing the range of potential minimum scalars across all inputs in an interpretable manner is not trivial. To generate the CRC in Table 2, we exploit the vector of lower bounds $\underline{k} = [k_1, k_2, ..., k_n]$ introduced by Theorem 4.2, which defines the minimum scalars required to generate specific feature value rules for a given translation. We select 5 equally spaced scalars from this vector (shown in Table 2 and in orange in Figure 6). The constrained selection of n scalars from $\underline{k}$, such that minimum costs are achieved over the inputs where recourse is found in Algorithm 1, can be viewed as a *monotonic submodular maximisation*. Problems of this type are NP-hard, though unlike AReS, our required search spaces $\underline{k}$ have shown empirically to be of significantly reduced size. Lastly, the coverage-cost and minimum costs properties of AReS are also shown in green for reference.

Coverage-Cost Profiles and Minimum Costs The numerical values in such a representation are easily portrayed via the *Coverage-Cost Profiles* and *Minimum Costs Histograms* (Figure 6, Upper and Lower, respectively). The figures shown demonstrate our framework's superiority over AReS (including efficiency being around 400 times higher in this situation), even in the context of categorical features, which AReS is designed for, favouring the use of the categorical translation theorems. As seen, these representations are particularly useful in comparing recourses between subgroups– optimal translations for one

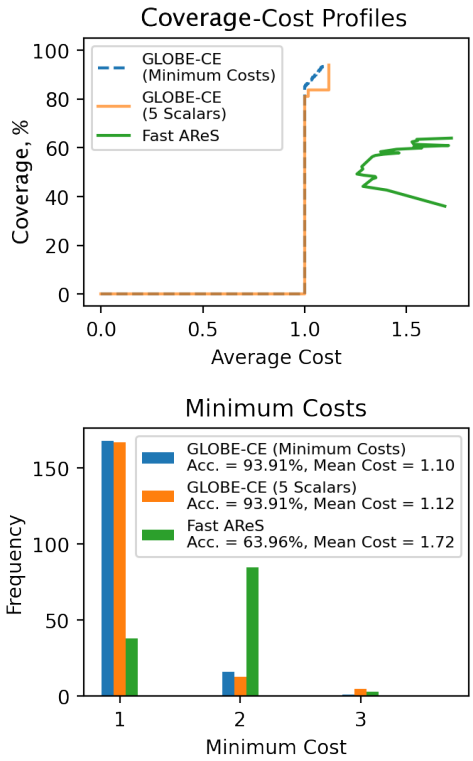

Figure 6. GLOBE-CE representations.

subgroup can also be directly applied to another, and vice versa, to gain insights as to a model's subgroup-level reasoning. Potentially, the rows of a CRC could too be reordered to achieve optimality. We discuss these future avenues for research in Appendix E.1.

Notes Regarding Scaled Translations By our definitions in Section 5 and Appendix E, the cost of a scaled translation $k\underline{\delta}$ on an input $\underline{x}_1$ increases monotonically with k, as displayed in Figure 7 (see Theorem 4.2 for proof w.r.t. categorical features). However, between inputs $\underline{x}_1$ and $\underline{x}_2$, a smaller scaled translation $k_1\underline{\delta}$ may induce a higher cost on input $\underline{x}_1$ than a larger scaled translation $k_2\underline{\delta}$ induces on input $\underline{x}_2$ (Figure 7). For a fixed $\underline{\delta}$, inputs $\underline{x}_1$ and $\underline{x}_2$ can have different costs. For example, should an 'If' condition be satisfied by $\underline{x}_1$ that is not satisfied by $\underline{x}_2$, and $\underline{x}_1$ and $\underline{x}_2$ are otherwise the same, then $\underline{x}_1$ will have a greater cost than $\underline{x}_2$ (Figure 7). These last statements apply if and only if categorical features are involved.

Scaled Translation Example (Single Categorical Feature) We conclude this section by providing one further example to aid understanding of the scaling process. Take a one-hot encoded feature with $n = 4$ values, translation $\underline{\delta} = [\delta_1, \delta_2, \delta_3, \delta_4] = [0, -1, 1, 1/2]$, index sequence of sorted δ (ascending) $\underline{\Delta} = [\Delta_1, \Delta_2, \Delta_3, \Delta_4] = [2, 1, 4, 3]$, vector of lower bound scalars $\underline{k} = [k_1, k_2, k_3, k_4] = [1, 1/2, \infty, 2]$. Recall that all rules will have the form 'If Feature Value = X, Then Feature Value = 3', since $\underline{\delta}_3$ is the maximum translation (Theorem 4.1). For example, $\underline{k}$ states intuitively that $\underline{\delta}$ must be scaled by a minimum factor of $k = 2$ in order for the translation to result in a rule for the fourth feature value, since $k_4 = 2$. If $k < 2$, inputs belonging to the fourth feature value will not flip value post-translation (take $k = 1$, where the post-translation value for such inputs is 1 for the third input and 1.5 for the fourth input, resulting in no change after re-encoding values by selecting the maximum). Table 9 illustrates the rule generation process in this case.

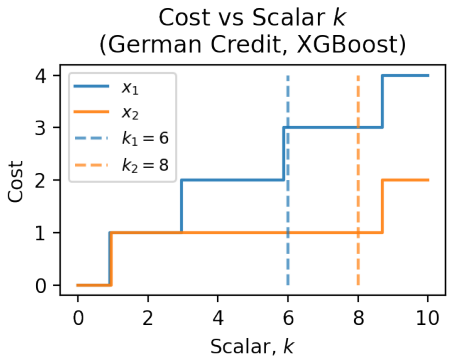

Figure 7. Display of a) monotonicity, b) smaller k induces higher cost, and c) fixed k induces variable cost.

	$m = 1$	$m = 2$	$m = 3$
Δ_m	2	1	4
k_{Δ_m}	1/2	1	2
$k_{\Delta_m} \times \delta_1$	0	0	0
$k_{\Delta_m} \times \delta_2$	-1/2	-1	-2
$k_{\Delta_m} \times \delta_3$	1/2	1	2
$k_{\Delta_m} \times \delta_4$	1/4	1/2	1
Rules for $k_{\Delta_m} < k \leq k_{\Delta_{m+1}}$	If 2, Then 3	If 1 or 2, Then 3	If 1 or 2 or 4, Then 3
Alternative Rules	If Not (1 or 3 or 4), Then 3	If Not (3 or 4), Then 3	If Not 3, Then 3

Table 9. Example of the vector of lower bound scalars $\underline{k}$ and its relation to a translation $\underline{\delta}$ and resulting rules. Values/indexes with distance $\geq$1 from the maximum translation ($k\delta_3$) are shown in red. $1 \leq m < n$.

B.3. Generality of our Framework

Though it has been mentioned, we will reinstate our objective of proposing a general CE generation framework, and proceed to qualify this goal with an in-depth analysis of the potential scope of our framework. This section delves deeper into the flexibility of GLOBE-CE to model specifics, over a variety of model families (DNNs, XGB, SVMs, results in Appendix D.3), as well as existing works in the model-agnostic domain of GCE research. We will conclude by considering the expansion of our implementation to current CE methods, and even to subsume the AReS algorithm.

Our motivation lies in demonstrating that our translation scaling techniques possess the requisite flexibility to be overlaid on top of existing criteria or algorithms in the CE space. Ideally, there should also exist specific parameter settings whereby our framework is equivalent to previous works, to guarantee at least equal performance. Analogously, Fast AReS is equivalent to AReS when using $\mathcal{RL}$*-Reduction*, $r = |V|$, and $s = |V|$. Any such reduction of r or s can be discarded if it hinders the performance of AReS, although we observe empirically that the opposite occurs. Similarly, the fact that our translation generation algorithm is completely general allows our framework to subsume any local method by simply setting the scalar $k = 1$ and the datapoints of interest to just the local input.

Model Specific – Deep Neural Networks DNNs are widely known for their backpropagation properties, which permit gradient calculations for the output layer w.r.t. the input. Several works thus utilise gradient descent to locate counterfactuals for such models. Since the goal of our general CE generation method $G(B, \mathcal{X}, n)$ is to locate n suitable GCE translations, we perform gradient descent on the following objective function, which takes both globality and categorical features into account:

$$\mathcal{L}(\underline{\delta}) = \frac{1}{|\mathcal{X}|} \sum_{\underline{x} \in \mathcal{X}} \left(B_0(\textbf{round}(\underline{x} + \underline{\delta})) + \lambda \textbf{cost}(\underline{x}, \textbf{round}(\underline{x} + \underline{\delta})) + \textbf{cat}(\underline{x} + \underline{\delta}) \right),$$

$$\text{where } \underline{\delta}_{\text{opt}} = \arg\min \mathcal{L}(\underline{\delta}).$$

In this context, B_0 is the softmax prediction of the undesired class, or equivalently, when moving beyond a recourse setting, the negative of such for the desired class. The **cost** and **round** terms are as previously described, and the additional **cat** term represents a penalty over categorical features in the resulting counterfactuals (we penalise categorical feature values that do not sum to 1). This naturally conforms to the standard framing of the counterfactual problem, and as previously, once δ_{opt} is converged to, it can be scaled to represent the final GCEs. For purely continuous datasets, the objective reduces to the minimisation of $\mathcal{L}(\underline{\delta}) = \frac{1}{|\mathcal{X}|} \sum_{\underline{x} \in \mathcal{X}} \left(B_0(\underline{x} + \underline{\delta}) + \lambda \textbf{cost}(\underline{\delta}) \right)$.

Model Specific – XGBoost Models These also provide an interesting direction for efficient CE generation, namely through readily accessible feature importance scores. In the case of a single decision tree, each node contributes a certain amount to the performance of the tree, and if each such contribution is weighted by the particular number of inputs that it influences, feature importances can be computed for the tree as a whole. Averaging the importances across all trees yields the overall importance per feature. We show a second time that model specific information, in this case feature importance, can be incorporated easily into our framework, simply by weighting the probability of each feature value in our random sampling framework with its feature importance.

Model Specific – Support Vector Machines We study here the linear kernel SVM, a model that offers mathematical expressions for the minimum distances or costs to the decision boundary. Knowledge of such is useful in assessing the performance of our framework; the minimum costs discovered by GLOBE-CE should be as close as possible to the theoretical minima provided.

In our problem landscape, we require minimum costs rather than minimum distances, and we proceed with two assumptions: costs are computed as the ℓ_2-norm of the individual feature costs; and features are continuous for the sake of this analysis. The first assumption is influenced by the fact that, in the context of linear kernel SVMs, ℓ_1 costs lead to completely sparse solutions (i.e. solutions where just one feature value changes). Though multiple optimal solutions do exist in this context which are not completely sparse, we wish to test the ability of both AReS and GLOBE-CE in finding a single, unique, and optimal GCE, which the minimum ℓ_2 cost translation uniquely provides.

We thus aim to minimise the cost $\|C\underline{\delta}\|_2$ of a translation $\underline{\delta}$. The feature costs vector is represented as a diagonal matrix C, scaling each feature independently before applying the ℓ_2-norm. Let the original input, counterfactual and translation be $\underline{x}_0$, $\underline{x}$, and $\underline{\delta} = \underline{x} - \underline{x}_0$, respectively. Given that the decision function of the SVM is $y_i = \underline{w}^T \underline{x}_i + b$, where $y = 0$ at the decision boundary, we derive:

$$y = \underline{w}^T \underline{x} + b = 0 \text{ and } y_0 = \underline{w}^T \underline{x}_0 + b$$
$$\implies -y_0 = \underline{w}^T (\underline{x} - \underline{x}_0) = \underline{w}^T \underline{\delta} = \underline{w}^T C^{-1} C \underline{\delta}$$

Recognising this expression as an inner-product between $\underline{w}^T C^{-1}$ and $C\underline{\delta}$, the Law of Cosines gives

$$-y_0 = \|\underline{w}^T C^{-1}\|_2 \|C\underline{\delta}\|_2 \cos\theta \implies \|C\underline{\delta}\|_2 = \frac{-y_0}{\|\underline{w}^T C^{-1}\|_2 \cos\theta},$$

which is minimised when $\cos\theta = 1$, such that the angle between $C\underline{\delta}$ and $C^{-1}\underline{w}$ is 0, or alternatively, $C\underline{\delta}$ is parallel to $C^{-1}\underline{w}$. This would imply that, upon normalisation, the two would be equivalent:

$$\frac{C\underline{\delta}}{\|C\underline{\delta}\|_2} = \frac{C^{-1}\underline{w}}{\|C^{-1}\underline{w}\|_2} \implies \underline{\delta} = \frac{-y_0}{\|\underline{w}^T C^{-1}\|_2} \times \frac{C^{-2}\underline{w}}{\|C^{-1}\underline{w}\|_2} = \frac{-y_0 C^{-2}\underline{w}}{\|C^{-1}\underline{w}\|_2^2}$$

Noting that $\underline{w}^T C^{-1} = C^{-1}\underline{w}$, since the cost matrix C is diagonal with $C = C^T$ and $C^{-1} = (C^{-1})^T$, we have thus derived the closed-form expressions for minimum cost $\|C\underline{\delta}\|_2$ and translation $\underline{\delta}$:

$$\|C\underline{\delta}\|_2 = \frac{-y_0}{\|\underline{w}^T C^{-1}\|_2} \text{ and } \underline{\delta} = \frac{-y_0 C^{-2}\underline{w}}{\|C^{-1}\underline{w}\|_2^2}$$

This translation on the local input ($\underline{x}_0$, y_0) in fact applies globally. We provide details regarding our user study in Appendix D.3 and show the primary results in Section 5.4 of the main text, demonstrating that the GLOBE-CE framework recovers minimum cost recourses very close to the theoretical global optima of the SVM.

Model Agnostic Other black box GCE methods that adopt translation based approaches, such as those in Plumb et al. (2020) and Ley et al. (2022), are easily integrated into our framework. Alternatively, we could view our framework as an extension to those works, filling the previous gaps with regards to minimum costs (by scaling translations) and categorical features (through our interpretation).

In reality, a plethora of issues surround past research on GCE translations. The algorithms in (Plumb et al., 2020; Ley et al., 2022) minimise *distance* between initial inputs (post-translation) and target inputs, resulting in a heavy reliance on the distribution of training data. If data lies too far from the decision boundary, GCEs learnt will not be well optimised for cost. In fact, any sparsely populated areas of the data manifold risk under-representation, despite possible significance w.r.t. the model's decision boundary. Furthermore, when the metric being optimised (the distance between datapoints) is not the metric used to evaluate the resulting explanations, tuning the learning process can become very difficult.

However, such problems can be bypassed with the introduction of our method to scale translations. Typically, the outputs of our random sampling framework are not as useful standalone, but embrace their full utility after the scaling process. Of course, our handling of categorical features would too be a great addition to these methods, which currently neglect to address categorical features.

The Role of Simplicity in our Framework The simplicity of our GCE generation process, as compared to earlier approaches (Rawal & Lakkaraju, 2020; Kanamori et al., 2022), serves to highlight the power of the consequent scaling and selection operations (outlined in this Appendix). It is particularly encouraging that we can achieve superior levels of reliability (high coverage, low cost) and efficiency compared to the baselines of AReS and Fast AReS, which holds promise for the scalability of GLOBE-CE; as datasets and models increase in complexity, there remains a large scope still for our algorithm to improve and adapt.

Amending our Framework As examples, genetic, evolutionary, or simplex (Nelder-Mead) search algorithms could improve both the reliability and the efficiency of the translation generation process as compared to random sampling. In fact, by framing the global search in an analogous sense to the local problem, as we do in Section 4, we leave the door open for any potential groundbreaking local CE research to be applied immediately in the context of GCEs. Another consideration is that a single translation direction will not always guarantee optimality. In fact, true optimality is only *guaranteed* by a single translation in contexts where the decision boundary spans a hyperplane (e.g. SVMs). We accommodate the use of multiple translations in our approach through a greedy maximum coverage search at the fixed cost of the search, but the presence of many works in this field outlining methods to achieve diversity should be noted, particularly given that our local framing of the global search directly permits the implementation of diversity techniques such as those in Mothilal et al. (2020).

Subsuming AReS We would further argue that the approach in AReS to produce a relatively large number of GCE explanations is one of the main contributors to its computational complexity, though the form of the two level recourse set generally requires as much. Of course, there is no reason that the apriori (Agrawal & Srikant, 1994) search could not also be used in our framework to generate translations for categorical features by influencing the probability of particular feature values in the same manner as the XGBoost feature importance method that we discussed earlier in this Appendix. On top of this, the ability of our framework to impose any number or combination of subgroup descriptors provides a route to essentially subsume the two level recourse set representation that AReS deploys, which we portray in Figure 1, with the exception (and advantage) that continuous features do not undergo binning.

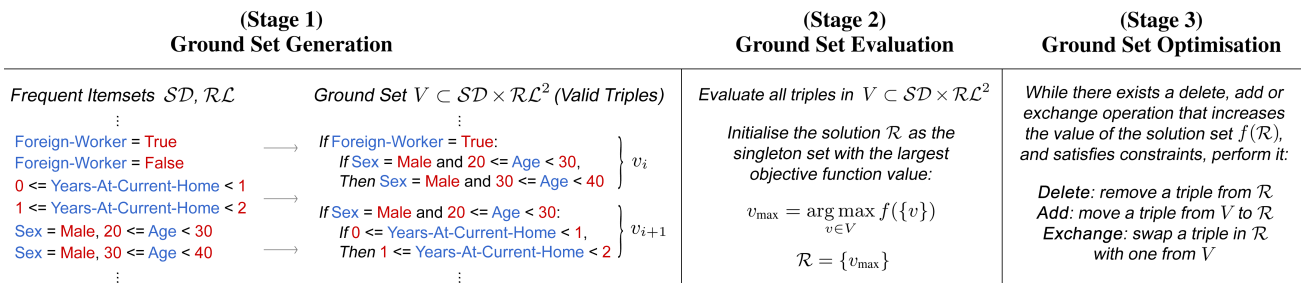

Figure 8. Workflow for our AReS implementation (bar improvements). Iteration over $\mathcal{SD} \times \mathcal{RL}^2$ computes all valid triples (Outer-If/Inner-If/Then conditions) in the ground set V (Stage 1). V is evaluated itemwise (Stage 2), and the optimisation (Lee et al., 2009) is applied (Stage 3), returning the smaller two level recourse set, R.

Comparing Subgroups The other notable detail in Figures 1 and 1 is the direct comparison of the same GCE between subgroups. Comparing GCEs with different feature values or relative scales can make it difficult to determine what, if any, biases are present in the recourses found. Conversely, by applying the same translation *direction* to both subgroups, the specific differences between recourses are much more easily identified, avoiding the metaphorical 'apples to oranges' comparison. We suggest that translations identified within subgroups are made *fluid*, and their reliabilities evaluated on opposing subgroups. Note that all of the representations that we propose in this text (coverage-cost profiles, minimum cost histograms, mean translations, and CRCs) support such translation fluidity.

C. AReS Implementation

We use this Appendix to provide further details regarding the implementation of each stage of the AReS workflow. Our implementation of AReS, without improvements, does in fact differ slightly from that proposed in Rawal & Lakkaraju (2020), and as such we will justify our changes herein. We acknowledge that this implementation is not guaranteed to be the most efficient possible, though hope that the patterns and improvements we have identified can aid further development of not only this framework, but others in the global counterfactual explanations space.

Given sufficient time, our AReS implementation produces results similar to those in the original paper, and the implementation only differs slightly (e.g., omitting the Bradley Terry model for cost, but retaining the same cost model between methods). Crucially, we implemented both AReS and GLOBE-CE as described in the algorithms of the respective papers, acknowledging that any further non-trivial optimisations might impact runtime. GLOBE-CE is simpler to implement, while AReS has more complex steps, which can affect efficiency depending on implementation (data types and CPU vs GPU).

AReS adopts an interpretable structure, termed two level recourse sets, comprising of triples of the form Outer-If/Inner-If/Then conditions (depicted in Figure 8). Subgroup descriptors $\mathcal{SD}$ (Outer-If conditions) and recourse rules $\mathcal{RL}$ (frequent itemsets output by apriori (Agrawal & Srikant, 1994)) determine such triples. Iteration over $\mathcal{SD} \times \mathcal{RL}^2$ yields the ground set V, before submodular maximization algorithm (Lee et al., 2009) computes the final set $R \subseteq V$ using objective $f(R)$. One strength of AReS is in assessing fairness via disparate impact of recourses, through user-defined $\mathcal{SD}$. While a novel method with an interpretable structure, AReS can fall short on two fronts:

Computational Efficiency Our analyses suggest that AReS is highly dependent on the cardinality of the ground set V, resulting in impractically large V to optimise. Our amendments efficiently generate denser, higher-performing V, unlocking the utility practitioners have expressed desire for.

Continuous Features Binning continuous data prior to GCE generation, as in AReS, struggles to trade speed with performance: too few bins results in unrealistic recourses; too many bins results in excessive computation. Our amendments demonstrate significant improvements on continuous data.

The overall search for a two level recourse set R can be partitioned into three stages, as detailed in Figure 8 and Table 10. We generate V, evaluate V, and optimise V, to return a smaller, interpretable set $R \subset \mathcal{SD} \times \mathcal{RL}^2$. As in the original work, we set $\mathcal{SD} = \mathcal{RL}$ and let $|\mathcal{RL}| = n \Rightarrow |V| < n^3$. Our optimisations include: *$\mathcal{RL}$-Reduction* and *Then-Generation*, which V faster or generate higher performing V, respectively (Stage 1); *V-Reduction*, a method which terminates evaluation of V after r or r' iterations, (Stage 2); and *V-Selection* which selects the best s items from the ground set (Stage 3).

Table 10. A summary of our AReS enhancements w.r.t. each stage of the search. We define α, n and $\mathcal{T}_i$ below.

	(Stage 1) Ground Set Generation	(Stage 2) Ground Set Evaluation	(Stage 3) Ground Set Optimisation
AReS	n^3 Iterations Performed	Evaluates Full Ground Set	Searches Full Ground Set
Fast AReS	$\alpha^2 n^3$ or $n^2 \max_i \mathcal{T}_i$ Iterations Performed	Evaluates and Reduces Full or Partial Ground Set	Searches Reduced and Sorted Ground Set

Of note is the scalability of AReS, which struggled with Default Credit and HELOC, datasets that contain significantly more points to explain ($|\mathcal{X}_{\text{aff}}|$) than German Credit or COMPAS, and significantly more continuous features. Additionally, the proportion of points with positive predictions (roughly 75% for German Credit and 45% for HELOC on average) influences the ease with which AReS finds recourses. For stringent models (those which scarcely predict positively), it would make sense that the vast majority of frequent itemsets generated by apriori are representative of feature value combinations that exist in the inputs with negative predictions, and we might therefore expect to need to generate an enormous number of triples before we can identify successful recourses.

Stage 1 Contribution ($\mathcal{RL}$-*Reduction*) We remove items with feature combinations that only occur once (e.g., "Sex = Male, Age < 30" has feature combination "Sex, Age"), yielding $|\mathcal{RL}| = \alpha n$ $(0 \leq \alpha \leq 1)$ in $\mathcal{O}(n)$ time. This generates an identical ground set V, yet saves $(1-\alpha^2)n^3$ iterations.

Stage 1 Contribution (*Then-Generation*, $\boldsymbol{q}$) At each iteration of $\mathcal{SD} \times \mathcal{RL}$, we filter the data by the If conditions and redeploy apriori to generate Then conditions. The ground set generated here differs from that in AReS, and we observe significant improvements on continuous features.

Stage 2 Contribution (*V-Reduction*, $\boldsymbol{r}, \boldsymbol{r'}$) At large $|V|$, evaluation is costly, yet this is a necessary requirement in finding high-performing triples. Fortunately, we can take advantage of two empirical observations: 1) the generation of a large ground set V is relatively cheap and 2) $acc(V)$[5] saturates far before the whole set has been evaluated. We evaluate a fixed number of triples and form a new ground set in one of two ways: adding each new triple; or adding only triples that increase the recourse coverage of the new ground set. We denote these r and r', respectively. For example, $r =$ None and $r' = 1000$ returns 1000 evaluations, and may store fewer than 1000 triples.

Stage 3 Contribution (*V-Selection*, $\boldsymbol{s}$) The bottleneck in the AReS framework, however, lies in the submodular maximisation of Stage 3 (Lee et al., 2009). We achieve speedups by further shrinking V pre-optimisation to a more practical starting point. We propose to sort the (new) ground set by recourse coverage (pre-computed) and select the s highest-performing triples. If $s = r$ or r', no sorting occurs.

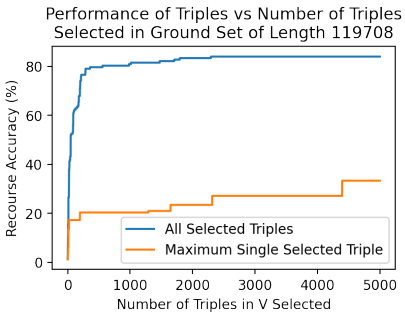

Figure 9. German Credit dataset. Redundancy in triples of ground set V. Note that *accuracy* in AReS is equivalent to *coverage* in our work.

C.1. Ground Set Generation (Stage 1)

AReS includes interpretability constraints for the total number of triples ϵ_1, the maximum width of any Outer-If/Inner-If combination ϵ_2 and the number of unique subgroup descriptors ϵ_3 in R. As in AReS, we take $\epsilon_1, \epsilon_2, \epsilon_3 = 20, 7, 10$. Constraints that are independent of the optimisation, such as ϵ_2, are applied in this stage in $\mathcal{O}(n^2)$ and not $\mathcal{O}(n^3)$ time. In our implementation, this involves expediting the ϵ_2 width constraint to the ground set generation process by constraining apriori to only return frequent itemsets that have length $\epsilon_2 - 1$ or less, since those already with width ϵ_2 cannot then be further combined with another itemset to form Outer-If/Inner-If conditions. If the width constraint is not violated for the If conditions, the resulting triple will automatically satisfy the constraint.

The implication of this is that we can apply the constraint in Stage 1, while we generate the ground set (in the first two levels of the iteration through $\mathcal{RL}^3$). This avoids applying the width constraint mid-optimisation in Stage 3, reducing the

[5]The *acc* function denotes recourse *accuracy* in AReS (equivalent to what we call *coverage* in this work).

time complexity of the operation from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$. It also reduces the number of constraints used in Lee et al. (2009), speeding up Stage 3. Since it makes sense that triples which violate the maximum width condition should not be generated in Stage 1, we assume that a similar approach is deployed (though not stated) in Rawal & Lakkaraju (2020).

Then-Generation The apriori algorithm (Agrawal & Srikant, 1994) alluded to in the main text takes a probability threshold p as input. This probability of an itemset in the data, or support threshold p, determines the size of $\mathcal{SD}$ and $\mathcal{RL}$, and consequently the size of V. Our *Then-Generation* method again utilises application of apriori, requiring a second support threshold q. A lower bound for the threshold q is derived here. In fact, there always exists a lower bound when mining frequent itemsets, such as in apriori, since no observed itemset can occur less than once. Thus, setting $q < 1/|\mathcal{X}|$ would be redundant. This allows us to analyse (in Appendix D.1) the effect of $1/|\mathcal{X} \leq q \leq 1$. Appendix C.4 further details apriori.

C.2. Ground Set Evaluation (Stage 2)

The submodular maximisation (Lee et al., 2009) first evaluates the objective function f over all triples $v \in V$, before initialising the solution R as the singleton set $\{v\}$ with the maximum $f(\{v\})$. For large $|V|$, this evaluation becomes computationally costly (more-so does the subsequent ground set optimisation), and many triples are also redundant. However, we require large $|V|$ in order to find high-performing triples and achieve an acceptable upper bound[6] on the final set, $R \subseteq V$.

Our improvement *V-Reduction* evaluates the objective function f (see Appendix C.3) over a fixed number of triples in V (recall that AReS evaluates the entirety of V). As we've demonstrated empirically, albeit on the four datasets tried in this investigation, evaluating the entire ground set is wasteful, given that performance of the first r elements of V saturates quickly, and more so if one considers that Stage 3 must then perform submodular maximisation over a space potentially hundreds of times as large, and that (Lee et al., 2009) only guarantees polynomial time.

However, there is a distinction between evaluating the objective function f and evaluating the acc and $cost$ terms used in evaluation. Fortunately, no extra major computation is required to evaluate the acc and $cost$ terms, since the objective function f returns model predictions and costs, and although the two processes differ, they can be carried out efficiently in tandem. This is promising, as not only does our method allow us to terminate evaluation once saturation has been reached, but it also provides us with the upper bound $acc(R) \leq acc(V)$. In many of our experiments, this upper bound is actually reached in Stage 3 far before the algorithm has completed, presenting us with a straightforward opportunity for early termination of the algorithm. This could further save time dramatically, and provide ease-of-use to practitioners, though was not included in our experiments.

C.3. Ground Set Optimisation (Stage 3)

We introduce two key modifications to Stage 3 of our implementation. The first is to the objective function, the second is to the submodular maximisation in (Lee et al., 2009).

Objective Function The objective function $f(R)$ in Rawal & Lakkaraju (2020) is designed to be non-normal, non-negative, non-monotone and submodular, and to have constraints that are matroids. These conditions are required for the submodular maximisation in Lee et al. (2009) to have a formal guarantee of convergence. This results in four terms in $f(R)$: *incorrectrecourse*, *cover*, *featurecost*, *featurechange*. Bar the *cover* term, all of these are subtracted from $f(R)$ (i.e., maximising correct recourse by maximising the negative of *incorrectrecourse*). Such an objective function with three adjustable hyperparameters can be very difficult to tune. For that reason, we also trial in our experiments an objective that consists very simply of $acc(R) - \lambda \times cost(R)$, which we maximise. We argue that the formal guarantees of convergence (polynomial time) are largely a misdirection of efforts in the original method. Polynomial time is not particularly helpful when the size of ground sets required for certain datasets/models is huge, and thus we instead focus on reducing the size of the ground set while retaining quality before the submodular maximisation (Lee et al., 2009) is applied.

Submodular Maximisation The algorithm states that, for k constraints, you can exchange up to k elements from your solution set R alongside the addition of one element from V. Stated also is that the optimisation should be repeated $k + 1$ times, before the best solution for R is then chosen. In reality, both of these induce high computational costs. Trivially, for the latter, ignoring the maximum width constraint (Appendix C.1) and taking $k + 1 = 3$, we will mostly increase the time taken by AReS three-fold. Having observed that both of these steps do not improve the performance of AReS significantly in

[6]For instance, if $acc(V) = 25\%$, we cannot achieve $acc(R) > 25\%$; conversely, a ground set with $acc(V) = 80\%$ requires major evaluation and will also include many low-performing, redundant triples.

our experiments, we omit them from the original and improved implementations. Furthermore, since the exchange operation stated is the most costly, our implementation checks for add/delete operations first until such options are exhausted. Note that works such as (Kanamori et al., 2022) instead utilise greedy algorithms, though still require in excess of three hours on simple datasets.

C.4. Apriori Interpretation

The apriori algorithm (Agrawal & Srikant, 1994) returns groups of itemsets that are frequently found within a dataset according to some support threshold (the probability of finding such an itemset in the data). Figure 10 demonstrates this (itemsets with less than or equal to 6 features). Rawal & Lakkaraju (2020) states:

1. A recourse set R is made up of triples of the form (d, c, c') denoting an outer if condition (d) and inner if-then conditions (c, c'), respectively (page 3).
2. "The corresponding features in c and c' should match" (page 3).
3. The optimisation searches for $R \subseteq \mathcal{SD} \times \mathcal{RL}$ (eq. 1, page 5).
4. "If the user does not provide any input, both $\mathcal{SD}$ and $\mathcal{RL}$ are assigned to the same candidate set generated by apriori" (page 5).

	support	itemsets
0	0.466387	(A11)
1	0.336134	(A12)
2	0.147059	(A13)
3	0.105042	(A32)
4	0.550420	(A33)
...	...	...
2042	0.126050	(A201, A192, A142, A91, A101, A174)
2043	0.117647	(A153, A201, A192, A91, A101, A174)
2044	0.105042	(A122, A201, A192, A142, A101, A174)
2045	0.109244	(A124, A153, A201, A142, A101, A174)
2046	0.172269	(A153, A201, A192, A142, A101, A174)

Figure 10. Apriori ($p = 0.1$).

We deduce that $\mathcal{SD}$ corresponds to the outer if conditions (d), so $\mathcal{RL}$ must then correspond to (c, c'). However, from statement 2, apriori cannot provide c, c' because of statement 2, which states that features must match, since a *single* apriori set is incapable of returning the 'Then' part of the recourse rule, yet statements 3 and 4 together imply that apriori generates the full space $R \subseteq \mathcal{SD} \times \mathcal{RL}$.

Furthermore, Lakkaraju et al. (2019) define the search space as $R \subseteq \mathcal{ND} \times \mathcal{DL} \times \mathcal{C}$, where $\mathcal{ND}$ and $\mathcal{DL}$ are analagous with $\mathcal{SD}$ and $\mathcal{RL}$, and $\mathcal{C}$ is the number of classes. The rules then take the intuitive form, with $\mathcal{ND}$, $\mathcal{DL}$ and $\mathcal{C}$ representing each part of the triple. Assuming this reasoning, alongside correspondence with the authors, confirms the form of our search space $R \subseteq \mathcal{SD} \times \mathcal{RL}$[2].

C.5. Further Discussion

Customising AReS As implicit in (Rawal & Lakkaraju, 2020), our implementation gives the user control over which features are dropped, how continuous features are binned, and the particular subgroup descriptors $\mathcal{SD}$ used for fairness analysis. We additionally posit that the particular data used to generate $\mathcal{RL}$ will potentially affect the final results quite significantly; there should be scope to assign the 'Then' conditions to an apriori evaluation on the dataset points with positive predictions, as these may be more likely to produce successful recourses. Data scarcity should be taken into account, as small datasets may not contain such a distribution of feature values that allow for effective searches of this nature. Finally, the constraint that all features in the Inner-If and Then conditions must match could possibly be ignored (the CET (Kanamori et al., 2022) and GLOBE-CE frameworks both take this approach).

Critiquing AReS We should preface this section by stating that the framework in (Rawal & Lakkaraju, 2020) is an original and major contribution to GCEs, and hope that its limitations can be overcome in consequent research. Save the shortcomings listed in the main text, we find a potential further three for consideration.

Firstly, AReS evaluates its recourses on the same set from which they are learnt (as do we in the main text). In practice, there are many scenarios in which we desire evaluation on a set of unseen test points. We perform such an evaluation, detailed in Appendix D, finding that performance does not deviate significantly, though the GLOBE-CE framework does generalise slightly better.

Secondly, the baselines used to assess AReS are notably weak (e.g. a naive averaging of instance-level explanations), and the effort exerted in tuning such baselines is unknown. While our work considers AReS to be state-of-the-art, owing to its performance against such baselines, making efforts to improve its performance (Fast AReS), perhaps a larger and more varied group of viable baseline frameworks for GCEs that strike a balance between naivety and sophistication could be conceived of.

Table 11. The keys **OG** (*Original AReS*), **RL** ($\mathcal{RL}$*-Reduction*) and **Then** (*Then-Generation)* refer to the generation process of the ground set, as per Appendix C.1. Arrows indicate values carried from one stage to the next. Apriori thresholds p and q are listed. Remaining parameters r, r' and s are listed in the Figure 12 plots.

	Stage 1	**Stage 2**	**Stage 3**
German Credit	**OG**: $0.169 \leq p \leq 0.390 \longrightarrow$ **RL**: $0.39 \leq p \leq 0.149 \longrightarrow$ **Then**: $0.9 \leq p \leq 0.303 \longrightarrow$ $q = 0.00125$	**OG**: $r = 5000$ **RL**: $r = 5000$ **Then**: $r = 5000, q = 0.00125$ **OG**: $0.316 \leq p \leq 0.26, r = \lvert V \rvert$	**OG**: $0.39 \leq p \leq 0.305, r = \lvert V \rvert$ **RL**: $p = 0.245$ **Then**: $p = 0.48,$ $q = 0.00125$
HELOC	**OG**: $0.325 \leq p \leq 0.285 \longrightarrow$ **RL**: $0.325 \leq p \leq 0.203 \longrightarrow$ **Then**: $0.75 \leq p \leq 0.563 \longrightarrow$ $q = 0.000127$	**OG**: $r = 5000$ **RL**: $r = 5000$ **Then**: $r = 5000, q = 0.000127$ **OG**: $0.325 \leq p \leq 0.3, r = \lvert V \rvert$	**OG**: $0.324 \leq p \leq 0.318, r = \lvert V \rvert$ **RL**: $p = 0.245$ **Then**: $p = 0.48,$ $q = 0.000127$

Finally, the computational expense of the method has not been documented to a strong degree, and the claims made regarding its formal guarantees and framework generality we find not entirely useful. The former claim, while potentially useful, only applies to the optimisation of the ground set V to produce the two level recourse set R, providing no concurrent guarantee on the upper bound of V, and pertaining also to polynomial time convergence, which scales particularly badly given the extremely high cardinality of V required to achieve acceptable performance. The latter claim, on the generality of the framework, neglects again to account for the size of the optimisation space required by AReS; in throwing a sufficient amount of compute at a simple dataset (far more than the original model required), one should reasonably expect to fit such exhaustive explanations.

Future Work on AReS Regarding the first point on generalisation above, we could extend our evaluation on unseen test data to include the effect of overfitting in models and out-of-distribution test points, given the susceptibility of current explanation methods to such inputs.

An interesting property of our Fast AReS optimisation is that there exist a variety of different shrunk ground sets, each with a high performance. While our optimisation simply picks one, multiplicity in AReS might be achieved by shuffling the ground set before the evaluation stage. We suggest the space of possible solutions be explored, tasked with answering the question: *where might this fail?*

Additionally, one might naturally question if a framework such as AReS could be extended beyond recourse, especially to other data forms such as images. While we explain the shortcomings AReS faces on continuous data, we posit that higher level views of image data such as latent spaces or concept embeddings could provide an interesting target for an extended future AReS framework.

Finally, as stated, we bin continuous features into 10 equally sized intervals. This follows the conventions in (Rawal & Lakkaraju, 2020), though we find that this approach struggles to trade performance with efficiency. Another facet of the framework is the direct interpretation of two bins; supposedly, one can move from any point in the first bin, to any point in the second bin, though this is not theoretically confirmed. The use of evenly spaced bins might also be improved upon in certain cases with quantile based discretisation, or more advanced decision tree structures. Lastly, the determination of costs is still a complex and unsolved problem in counterfactual literature– we justify our approach in Appendix E.

D. Experimental Results

This Appendix further details various experiments on AReS, Fast AReS and our GLOBE-CE framework. User studies, model specific analyses, and hyperparameters are included.

D.1. AReS Optimisations

As AReS struggles to achieve sufficient coverage within reasonable times, we set hyperparameters for *featurecost* and *featurechange*, or λ, to 0, also finding that the average costs were low and did not vary a large amount, justifying the decision to target correctness. The remaining hyperparameters used in the Figure 12 experiments are as detailed per stage in Table 11. Recall also that we have bounded the range of the apriori threshold q used in *Then-Generation* to $1/|\mathcal{X}| \leq q \leq 1$ (Appendix C.1). Figure 11 demonstrates that for $q > 1/|\mathcal{X}|$, we slightly reduce the runtime, at the expense of a much larger drop in performance. Observe that the red and brown lines (where p is held constant and q is varied) converge to the green and purple lines (where $q = 1/|\mathcal{X}|$ and p is varied), respectively. The brown and purple plots also indicate that combining

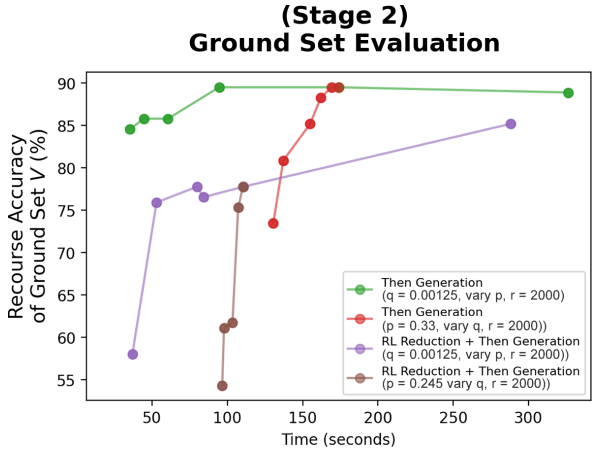

Figure 11. Effect of apriori threshold q in the proposed *Then-Generation* method (German Credit). As noted previously, recourse *accuracy* in AReS is equivalent to *coverage* in our work.

our two improvements $\mathcal{RL}$*-Reduction* and *Then-Generation* performs sub-optimally. We thus decide to evaluate these improvements separately, with a fixed $q = 1/|\mathcal{X}|$ threshold. We note that the choice of $\mathcal{SD} = \mathcal{RL}$ weakens performance, which aids in stress-testing scalability.

D.2. Fast AReS Experimental Improvements

The evaluation of Fast AReS is broken down per workflow stage. For various hyperparameter combinations (r, r' and s), the final sets returned in Stage 3 achieve significantly higher performance within a time frame of 300 seconds, achieving accuracies for which AReS required over 18 hours on HELOC. Appendix D lists details in full, including the combinations of hyperparameters used.

Reliability & Efficiency In Stage 1, we demonstrate $\mathcal{RL}$*-Reduction* is capable of generating identical ground sets orders of magnitude faster, and *Then Generation* constructs (different) ground sets rapidly. Stage 2 shrinking ($r = 5000$) significantly outperforms full evaluation, and *Then Generation* erases many continuous data limitations by short-cutting the generation of relevant rules. Finally, Stage 3 demonstrates vast speedups, owing to the generation of very small yet high-performing ground sets in the previous stage: r, r' and s restrict the size of V yet retain a near-optimal $acc(V)$.

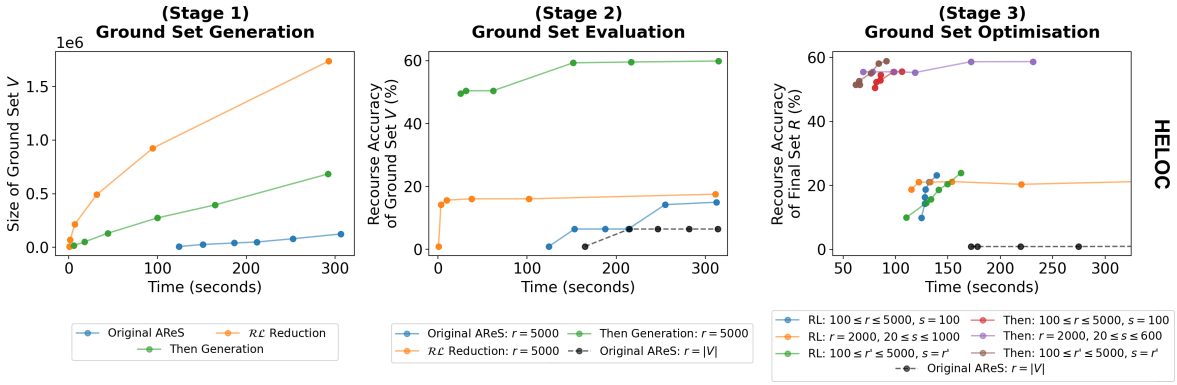

Figure 12. Fast AReS improvements (HELOC). Left: Size of ground set V vs time. Centre: Accuracy of V vs time. Right: Final set $acc(R)$ vs time. For other *categorical* datasets, $\mathcal{RL}$*-Reduction* achieved speedups similar to *Then-Generation*. For Default Credit (predominantly continuous), the latter performed best. Again, *accuracy* in AReS refers to *coverage* under our definitions.

D.3. GLOBE-CE

Model Specific Analyses The GLOBE-CE implementations that utilised some sort of model information (DNN gradients, XGB feature importance scores) are depicted in Figures 13 and 14, alongside GLOBE-CE (blue), diverse GLOBE-CE with $n = 3$ (orange) and Fast AReS (green).

The use of **DNN gradient descent** to determine an appropriate translation prior to scaling, shown in red in Figure 13, demonstrates marginally improved performance over GLOBE-CE on continuous data (HELOC, Right), though the same method for categorical data (German Credit, Left) produces similar results to GLOBE-CE, with performance dependent on the particular coverage to cost trade-off desired (as expected, given that gradient descent can struggle in the presence of categorical features).

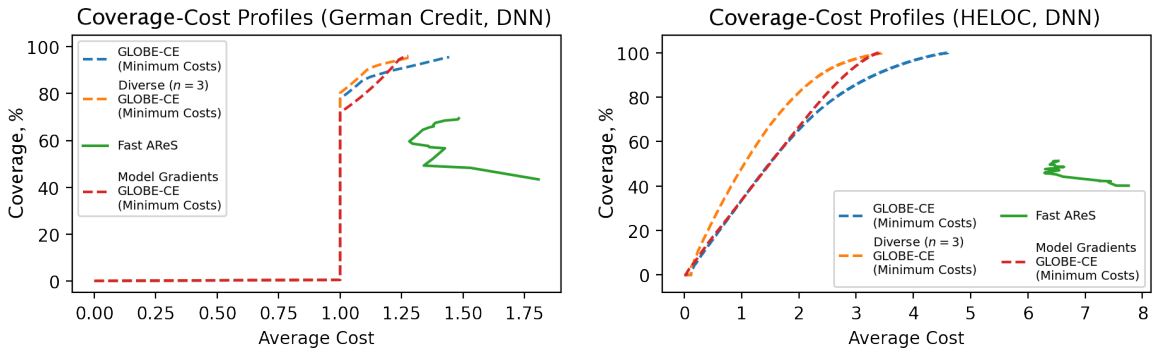

Figure 13. Translations generated by gradient descent and scaled (red), compared to the other frameworks.

The use of **XGB feature importances** to weight the probability of the random translation generated prior to scaling, shown in red in Figure 14, demonstrates improved performance over GLOBE-CE on continuous data (Default Credit, Right) and, marginally, on categorical data (COMPAS, Left).

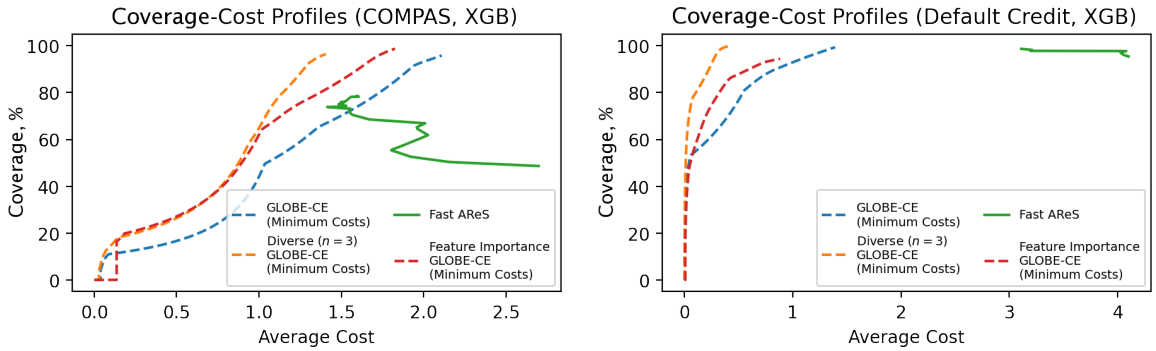

Figure 14. Translations generated by feature importance and scaled (red), compared to the other frameworks.

The results across these two particular model classes illustrate the flexibility of our framework and its ability to accommodate model specific properties, doing so effectively with model gradients (DNNs) and feature importance scores (XGB). The effectiveness of the random sampling framework in the absence of any additional model information other than its predictions is also demonstrated.

D.4. User Studies

In our experiments, participants received either explanations from GLOBE-CE or from AReS, and were subsequently asked to provide bias identification, and a description if they believed a bias to exist. The aim of the study is two-fold. We first hope to establish the importance of providing the underlying distribution of costs as per the GCEs output by a framework (i.e., the proportion of inputs that correspond to each GCE, given that the lowest cost GCE that results in a flipped prediction

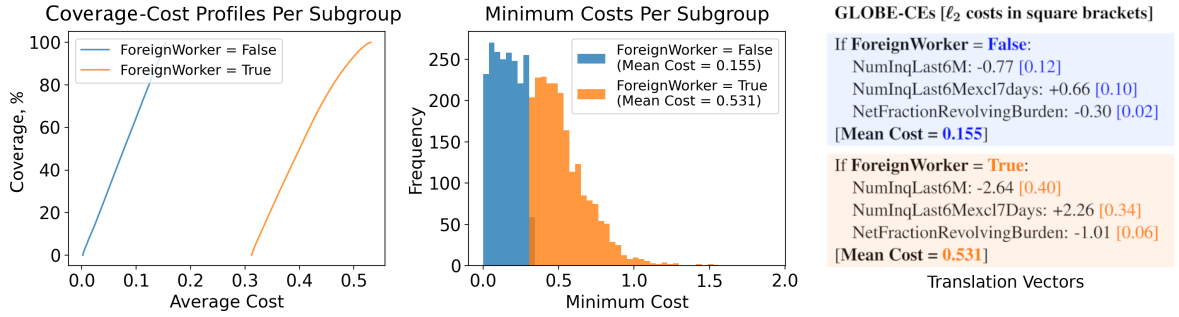

Figure 15. Black Box 2 user study (GLOBE-CE). Example comparisons with synthetic ForeignWorker subgroups. Left: Coverage to cost trade-offs (covering more inputs requires larger magnitudes) Centre: Minimum costs per input. Right: Mean translation direction for each subgroup. Mean cost is computed as the ℓ_2 norm of the mean translation after features are weighted.

is selected for each input). Secondly, we hoped to reinforce our claim in Section 3 that sub-optimal recourse costs/accuracies can resulting in misleading conclusions regarding bias.

In the first question of the study, both frameworks output the exact same explanations, minus the fact that GLOBE-CE also includes costs and the percentage of inputs for which a particular rule was best. Explanations for AReS/GLOBE-CEs are identical (minus [*square brackets*] for AReS). This is shown below in the Cumulative Rules Chart, where the subgroup *Female* is discriminated against by the model, though due to the data distribution, the subgroup *Male* is discriminated against with respects to the costs required for recourse. This challenges the assumption made in AReS that recourse biases can be simply gauged without knowledge of the inputs affected.

Cumulative Rules Chart:

If Sex = Male:

If Job = No **and** Property = No,
Then Job = Yes **and** Property = Yes } Rule M1 [*Cost 2, 90% of Inputs*]

If Healthcare = No,
Then Healthcare = Yes } Rule M2 [Cost 1, *10% of Inputs*]

If Sex = Female:

If Job = No **and** Property = No **and** Savings = No,
Then Job = Yes **and** Property = Yes **and** Savings = Yes } Rule F1 [*Cost 3, 10% of Inputs*]

If Healthcare = No,
Then Healthcare = Yes } Rule F2 [*Cost 1, 90% of Inputs*]

In the second question of the study, we introduce a synthetic subgroup *ForeignWorker* to HELOC, and discriminate against it by forcing recourse costs to be higher. We use a linear kernel SVM as our model, in order to compare the recourse costs of AReS and GLOBE-CE against the absolute minimum costs, which we provide a theoretical analysis for in B.3. Our results for GLOBE-CE are as in Figure 15, achieving near-optimal costs, and resulting in correct bias description by users.

E. Discussion & Future Work

Having outlined a highly flexible framework for GCEs, it is important that we consider now both the possible limitations of the GLOBE-CE framework, as well as avenues for future work or growth.

E.1. Possible GLOBE-CE Limitations and Future Work

Determining Costs of Actionable Recourse We do not claim to hold the answers regarding the costs associated with actioning particular counterfactuals. In reality, this is an incredibly complex problem to solve. Not only could the costs of certain actions vary with time, depend on each other and themselves, or be susceptible to unknown degrees of randomness, but they could also depend on a multitude of factors surrounding the specific individual that is ultimately tasked with the

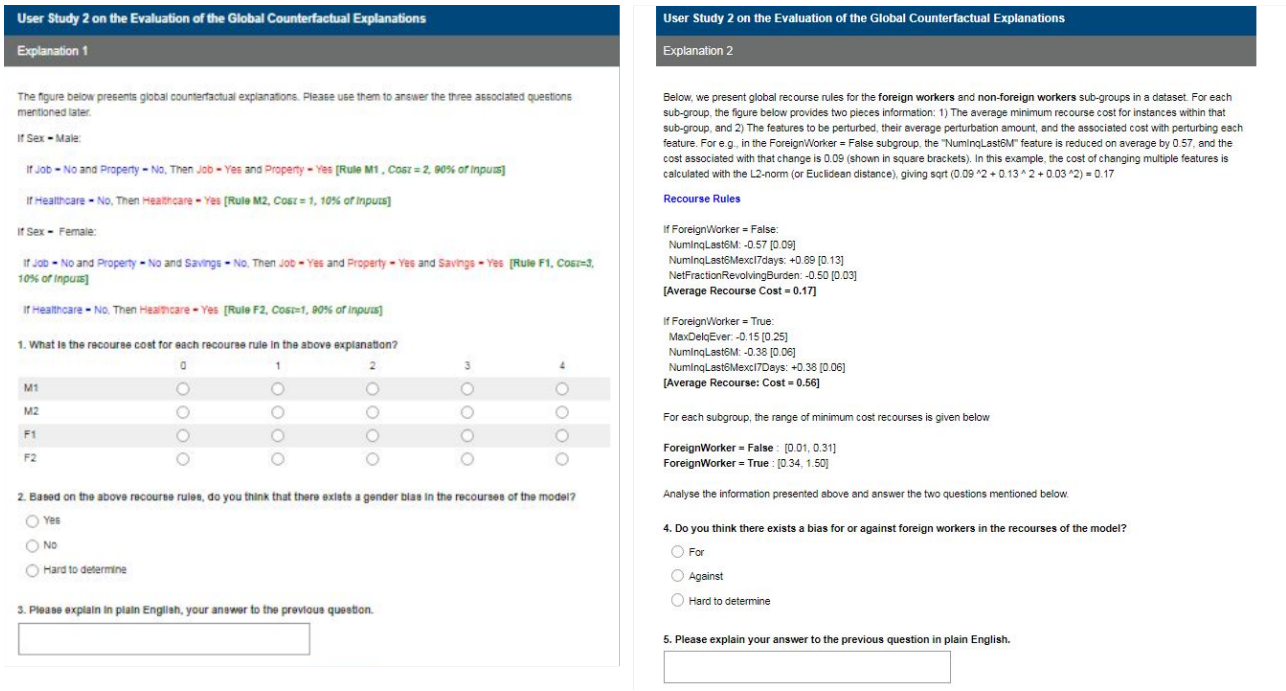

Figure 16. User study 2 snapshots (GLOBE-CE). Left: Explanation 1. Right: Explanation 2. Successful recourse bias identification and description occurs when underlying costs distributions are shown.

action. Disregarding such difficulties, building a model that perfectly reflects a user's struggle in executing a particular set of actions might easily require an improper breach of said user's privacy.

Rawal & Lakkaraju (2020) proposes the use of the Bradley-Terry model to compute fixed feature costs, on the basis that pairwise feature comparisons are relatively easy for experts to make, though this does neglect to account for the properties of the end user or even the dependencies between costs. Notwithstanding this, searching for *better* cost models remains a worthwhile area of research. For instance, the use of the Bradley-Terry model would surpass in performance the use of non-normalised ℓ_1 distance as cost. In light of all of this, we do not overly focus on cost estimation in this paper, acknowledging that actionability remains, to all intents and purposes, an unsolved problem, and recognising that this could affect the reliability of bias assessment between subgroups based upon costs. Instead, we settle for the use of unit costs between categorical features or per decile of continuous features (ℓ_1).

Expanding our Baselines The baselines used in AReS are discussed in greater detail in Appendix C.5. Further possible candidates are the GCE translations proposed in (Plumb et al., 2020; Ley et al., 2022), and a non-interpretable accumulation of the *costs* of local CEs, used solely to assess minimum costs per input, and not naively averaged over to produce GCEs. While this last suggestion doesn't yield GCEs, it could offer a challenging set of minimum costs per input for our framework to seek to outperform.

Limitations of Categorical Translations Our form of categorical translations always yield rules of the form "If A (or B or C ...), Then X" for any one particular feature (with possible negation on the "If" term). However, it may also be useful to represent other forms, such as "If A, Then Not A" or "If A then (B or C)". The minimum costs yielded would not decrease as a result of these new forms, since they can still be represented with the original (e.g. "If A, Then B" would solve simply the first suggestion), though may have interpretability implications which could be explored in future work.

Alternative GCE Approaches The common perspective is to discover CEs from a set of inputs, though a user-based approach which does the opposite could also be proposed, whereby a user specifies a CE of interest, and the group of inputs most strongly affected by such a change are identified and returned. Such an approach could utilise our proposed scaling operation to better summarise the group of inputs in terms of both coverage and minimum cost, and could also execute without the need for user-specified CEs, generating, scaling and analysing CEs automatically instead.

Relation between CEs and GCEs We postulate the thought experiment that, for a given dataset, there could always conceivably exist further inputs or dimensions at larger or smaller scales whereby global explanations are reduced to local explanations, and vice-versa. For example, a local CE, where all input dimensions are accounted for, can suddenly become a subgroup explanation if extra dimensions are added to the dataset. By the same logic, removing dimensions or adding datapoints can render global explanations local. In the context of fixed models that we find ourselves in here, such modifications possess more theoretical than practical potential, though we pose that such a direction for future research could uncover interesting properties between local and global CEs.

GLOBE-CE: Extension We *translate* N negatively predicted instances to N neutrally predicted instances (on the decision boundary). This paves the way for further optimisation of the translation. For example, if the nearest neighbours between the original points and those on the decision boundary do not correspond to the translation for a significant number of inputs, further tuning could ensue.

E.2. Robustness of Counterfactuals

Recent research has demonstrated instabilities w.r.t. CE techniques from state-of-the-art methods:

- CEs are non-robust (become invalid) in certain scenarios. (Dominguez-Olmedo et al., 2021; Mishra et al., 2021; Dutta et al., 2022)
- Minor perturbations to input features may result in substantially different recourses from popular CE approaches. (Slack et al., 2021)
- Minor changes in the underlying ML model, for example, due to retraining with new data, makes CEs for the old model invalid on the new model. (Upadhyay et al., 2021)
- Noisy human implementation of the suggested recourse may prevent an individual from achieving the desired response from an ML model. (Pawelczyk et al., 2022)

Related research has also identified a positive link between the distance moved past a decision boundary and counterfactual robustness, which suggests that GLOBE-CE can easily be modified to account for robustness by simply increasing the scalar k that is applied to a global translations δ.