

Do Personalized Evaluation Functions Reflect Human Preferences? A Study of Weighted Proximity in Algorithmic Recourse

Firdaus Ahmed Choudhury, Ethan Leicht, Jude Ethan Bislig, Arsh Gupta,
Akshaya Shyamsundar Rekha, Shubham Gupta, Haren Anand, Hangzhi Guo, Amulya Yadav

The Pennsylvania State University
University Park, PA 16802 USA

Firdaus@psu.edu, ebl5321@psu.edu, jeb6596@psu.edu, abg6210@psu.edu, ajs9686@psu.edu,
skg5962@psu.edu, hqa5367@psu.edu, Hangz@psu.edu, Amulya@psu.edu

Abstract

Algorithmic recourse focuses on providing individuals (end-users) negatively affected by ML-based decisions with recourse (actionable recommendations) that help them achieve a more favorable outcome in the future. Within this field, counterfactual explanations (CFEs) have gained popularity as a means of providing actionable recourse to end-users by identifying the minimal changes in their feature values needed to flip the ML model’s prediction to a more desirable one. Almost all prior work in CFEs rests on the assumption that minimizing (weighted) proximity (i.e., weighted L_1 norm between the original input and the CFE) is a good proxy metric for capturing real-world human preferences with regard to the evaluation of prospective CFEs. Unfortunately, no prior work has empirically evaluated the effectiveness of weighted proximity in capturing real-world user preferences for evaluating CFEs. To address this research gap, this paper makes three novel contributions. First, we develop a novel two-phase study design that helps us effectively learn feature-specific weights for each individual end-user (in Phase 1) and evaluate the effectiveness of weighted proximity using those learned user-specific weights (in Phase 2). Second, we conduct the first-ever set of user studies (with 101 participants) that focus on evaluating the efficacy of *global* and *personalized* weighted proximity in capturing user preferences. Our experimental results show that global weighted proximity aligns with user preferences in only half the cases (53.21%), and personalized weighted proximity is not consistently better (56.91%). Finally, we conduct a series of sensitivity analyses to confirm that our findings are robust across different experimental settings and modeling choices.

1 Introduction

Machine Learning (ML) models are increasingly being deployed in domains that directly affect human lives (Huang et al. 2023; Zhang et al. 2023; Verma et al. 2024b). With this rapid adoption, there is a growing demand not only from end users (Binns et al. 2018) but also from regulatory frameworks such as the EU’s General Data Protection Regulation (GDPR) for greater accountability and transparency in AI/ML-based decision-making systems (Goodman and Flaxman 2017). This demand has led to a surge in research on Explainable AI (XAI), a field focusing on developing

methods to make the decision-making process of complex ML models more interpretable and transparent (Mersha et al. 2024).

Within this extensive body of research, counterfactual explanations (CFEs) have emerged as a promising method for providing intuitive and actionable insights into the decision-making processes of black-box ML models (Dodge et al. 2019; Verma et al. 2024a; Guo et al. 2024). CFEs provide post-hoc transparency by identifying the minimal changes in feature values needed to convert an input point x into x' , a new artificially constructed data point (called the CFE) that results in a different (more favorable) prediction outcome than the original data point x (Wachter, Mittelstadt, and Russell 2018). For instance, if a loan applicant is denied credit by a bank’s ML model, the corresponding CFE might suggest that increasing the applicant’s income by \$2,000 would result in an approval.

Existing algorithms for generating CFEs are formulated as optimization problems that minimize “cost of change” (i.e., the amount of modification required in changing x to x') as one of the key objectives (Wachter, Mittelstadt, and Russell 2018; Brughmans, Leyman, and Martens 2023). Almost all prior work in this space operationalizes this notion of “cost of change” as *proximity* or *sparsity* objectives (defined as the L_1 and L_0 norm distance between x and x' , respectively) (Guo, Nguyen, and Yadav 2023). To date, despite scant empirical evidence, both proximity and sparsity metrics have been assumed to be good proxies for capturing human preferences for desiring minimum changes in CFEs (Ustun, Spangher, and Liu 2019). However, recent work has shown through user studies that CFEs optimized for proximity (or sparsity) do not align well with people’s preferences (Tominaga, Yamashita, and Kurashima 2024). Independently, several researchers have pointed to the possibility of using weighted proximity (i.e., the weighted L_1 norm between x and x' , where the feature-specific weights represent the difficulties faced by users in modifying that feature) as an alternative metric that should, in theory, better capture people’s preferences (Assaad et al. 2021; Mc Grath et al. 2018).

To the best of our knowledge, no attempt has been made to validate the effectiveness of weighted proximity in capturing people’s preferences. For example, Rawal and Lakkaraju (2024) propose a framework for learning personalized fea-

ture weights to model user-specific difficulty and using those to generate tailored recourses¹. However, they evaluate the efficacy of their learning method only on simulated data.

This lack of rigorous human validation highlights a fundamental gap in counterfactual explanations research: without testing evaluation functions (such as weighted proximity) with real users, we cannot know whether they truly reflect the needs of the end user. As a result, the disconnect between how researchers evaluate CFEs and how users interpret or act on them continues to grow, potentially leading to solutions that are technically sound but practically inapplicable.

To address this research gap, this paper makes the following three contributions. First, we develop a novel two-phase user study design that helps us effectively learn feature-specific weights for each participant in Phase 1, and evaluates the effectiveness of weighted proximity using the learned weights in Phase 2. Second, we conduct the first empirical evaluation of global and personalized weighted proximity for capturing user preferences in CFE generation, involving 101 participants across a pilot study and two independent user studies in a loan rejection setting. Our experimental results show that global weighted proximity aligns with user preferences in only 53.21% of cases, and personalized weighted proximity yields only modest and inconsistent alignment of 56.91% across the two studies. Finally, we conduct a series of sensitivity analyses that confirm the robustness of our results across different experimental settings and modeling choices. This work underscores the need for CFE research to be grounded in empirically validated user-centric evaluations, rather than solely on metrics that are mathematically motivated but empirically unvalidated.

2 Related Work

2.1 Counterfactual Explanations

Within the XAI community, CFEs have gained recognition for their ability to illustrate how black-box ML models arrive at their predictions (Byrne 2019; Verma et al. 2024a). As CFEs identify the smallest set of changes needed to alter a prediction, they naturally align with the goals of algorithmic recourse: to help users understand the cause behind unfavorable decisions they received from a predictive model and subsequently provide actionable steps they can take to overturn those decisions (Ustun, Spangher, and Liu 2019; Tominaga, Yamashita, and Kurashima 2024).

Prior literature has proposed a wide range of approaches to generate CFEs, differing in their optimization objectives, generation mechanisms, and the accessibility of the underlying prediction model (Verma et al. 2024a). Some of the most common approaches for generating CFEs include gradient-based optimization methods (Wachter, Mittelstadt, and Russell 2018; Schut et al. 2021), decision-tree-based methods (Fernández et al. 2020; Carreira-Perpiñán and Hada 2021), mixed-integer programming formulations (Karimi et al. 2020; Ustun, Spangher, and Liu 2019), generative-model-based approaches (Mahajan, Tan, and Sharma 2019),

and search-based methods such as DiCE (Mothilal, Sharma, and Tan 2020). Despite differences in generation methods and constraints, most approaches rely heavily on proximity (L_1 norm distance) and sparsity-based (L_0 norm distance) objectives to evaluate and generate CFEs.

For example, Wachter, Mittelstadt, and Russell (2018) introduced an optimization-based method that jointly optimizes CFEs on a combined loss of validity and proximity via gradient descent. Subsequent methods have built on their work and emerged with additional objectives (in addition to proximity or sparsity), such as diversity (Mothilal, Sharma, and Tan 2020), and feasibility (Mahajan, Tan, and Sharma 2019). Recently, however, Tominaga, Yamashita, and Kurashima (2024) have provided empirical evidence that CFEs optimized for proximity (or sparsity) do not match user-preferred recourses, highlighting the ineffectiveness of proximity (and sparsity) as user-centric metrics for CFE evaluation.

2.2 Weighted Proximity in CFEs

To address the limitations of standard proximity metrics, several works have proposed incorporating feature-specific costs to better reflect the difficulty of modifying certain features (Yadav, Hase, and Bansal 2022; Rawal and Lakkaraju 2020). These approaches assign different weights to features, enabling the optimization process to penalize changes to specific attributes more heavily than others (Mc Grath et al. 2018; Assaad et al. 2021). Formally, **weighted proximity** assumes a distinct feature weight w_i for feature i (that reflects its perceived cost of modification to a user), and is defined as the weighted L_1 norm between input point x and CFE x' .

$$\text{WeightedProx}(x, x') = \sum_{i=1}^d w_i \cdot \frac{|x_i - x'_i|}{\text{Norm}_i}$$

A common assumption in prior CFE literature on weighted proximity is that the perceived difficulty of modifying a feature is often assumed to be uniform across all users. This presumption leads to a set of global feature weights w_i^g (for feature i), where a single set of shared feature weights is used for all individuals in a population. The weighted proximity that is calculated using global feature weights w_i^g is called *global weighted proximity*. Numerous CFE methods implicitly adopt this formulation by incorporating a single set of feature weights/costs as a parameter that is used for all recourse generations (Wachter, Mittelstadt, and Russell 2018; Mothilal, Sharma, and Tan 2020; Liu et al. 2025).

Recent work extends this formulation by learning personalized feature weights w_i^a for each user a . For example, Rawal and Lakkaraju (2024) propose a framework for inferring *personalized* feature-specific weights w_i^a using pairwise comparisons via the Bradley-Terry model (Bradley and Terry 1952). Their method enables recourse costs to vary across users based on individual preferences; however, their approach was tested only on simulated data. To the best of our knowledge, no empirical evaluation has assessed the efficacy of (global and personalized) weighted proximity as a

¹Counterfactual explanations and algorithmic recourse are used interchangeably in the rest of the paper

proxy for user preferences in real-world settings. Our work tackles this challenge and provides the first empirical validation of global and personalized feature-weighted proximity for CFE generation with real human users.

2.3 User Evaluation of Algorithmic Recourses

There is a growing body of research emphasizing the importance of empirically validating CFEs with real users. Keane et al. (2021) critically examined gaps in the evaluation of counterfactual methods and highlighted that only 21% of studies included user evaluations of specific counterfactual methods, and only 7% conducted comparative user studies to benchmark different approaches. Another shortcoming in the counterfactual literature that they underscored was the lack of comparative testing across methods; they noted that only 40% of papers include any form of comparative studies. In response to this gap, recent studies have focused on user-centered evaluations in counterfactual research.

Tominaga, Yamashita, and Kurashima (2024) conducted an empirical user study involving 362 participants, demonstrating that CFEs optimized for technical objectives, such as proximity and sparsity, often misalign with user preferences. These findings reinforce the argument that strong results on simulated benchmarks do not necessarily translate to acceptable recommendations in real-world contexts. Upadhyay, Lakkaraju, and Gajos (2025) performed multiple user studies with approximately 360 participants, comparing CFEs with other recourse recommendation strategies. Their results showed that simpler feature-based recommendations outperformed CFEs. Collectively, these findings underscore the need for human-centered evaluations to meaningfully assess the utility of CF generation methods. Our work is motivated by the need to conduct user evaluations that not only validate counterfactual methods but also provide a richer understanding of how users reason about and choose between alternative recourses.

3 Preliminaries

For completeness, we describe the pairwise preference learning methods from prior work that are used in this paper to infer feature-specific weights.

3.1 Pairwise Preference Learning Models

Preference learning models are designed to infer latent utility functions based on observed human choices (De Peuter et al. 2024). A common approach within these frameworks is to infer preferences through pairwise comparisons, in which a probability function is estimated by analyzing repeated selections between two alternatives (Bradley and Terry 1952; Thurstone 1927).

In the context of CFEs, Rawal and Lakkaraju (2024) propose inferring feature-specific weights, w_i , using a Bradley-Terry (BT) model. The BT model is a probabilistic framework that estimates a latent strength parameter β_i for each feature i by modeling the likelihood of one feature being preferred over another (Bradley and Terry 1952). The probability that modifying feature i is preferred over feature j is

modeled as a logistic function:

$$P(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

Additionally, we consider the Thurstone-Mosteller (TM) model (Thurstone 1927; Mosteller 1951) as an alternative framework for learning feature preferences. Unlike the BT framework, which models pairwise preferences using a logistic function, the TM model assumes an underlying Gaussian utility distribution over feature preferences and estimates pairwise preference probabilities using the cumulative distribution function (Maystre and Grossglauser 2011):

$$P(i > j) = \Phi\left(\frac{\beta_i - \beta_j}{\sigma}\right) \Big|_{\sigma=1} = \Phi(\beta_i - \beta_j)$$

where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution, and σ is the variance of the underlying utilities.

Given pairwise choices between alternative CFEs, preference models estimate a latent strength parameter β_i for each feature i , where a larger β value reflects the increased desirability of modifying that feature. Since a recourse recommendation consists of a set of feature modifications, and pairwise preference models primarily operate on feature-level comparisons, recourse-level choices must first be decomposed into individual pairwise feature comparisons. Given two alternative recourses being compared, each feature uniquely modified in the preferred recourse is treated as being preferred over every individual feature modified in the alternative recourse. These pairwise feature comparisons are aggregated into a wins matrix, which the preference model uses to infer feature-specific β values. The learned strength parameters β_i are transformed into feature weights by taking their additive inverse ($w_i = -\beta_i$), such that features perceived as less preferable to modify (smaller β_i values) receive larger weights.

4 Methodology

This section outlines our methodology for generating and evaluating CF recourses, which serves as the basis for the design and implementation of our user studies. Consistent with prior work in algorithmic recourse, we study recourse generation in the context of loan rejection, where participants receive recommendations on modifying their applications to obtain loan approval. In our setting, a CFE is a set of minimal feature changes to the original loan application that would result in the loan being approved.

This section is organized as follows. First, we describe the different datasets, prediction models, and CFE algorithm choices used in each of the two studies. These components are used to construct two datasets of “*CFE scenarios*”, where each *CFE scenario* consists of a rejected loan application paired with two alternative CFEs (we use this terminology of *CFE scenarios* consistently throughout this paper). We then provide details of our participant recruitment procedure, followed by the shared two-day study design used across both user studies.

4.1 Study Overview

To evaluate whether weighted proximity aligns with user preferences, we conducted two independent user studies. Consistent with prior work (Tominaga, Yamashita, and Kurashima 2024; Rawal and Lakkaraju 2024), the evaluation was performed in the loan rejection domain, which is one of the most widely studied contexts in algorithmic recourse literature.

The first study was conducted in a controlled environment using a synthetic dataset with a decision tree-based loan prediction model, in which candidate CFEs were generated by traversing the decision tree (details in Section 4.3). To corroborate our findings from the first study, a second study was conducted with a different participant group in a more realistic setting. In this study, we utilized a publicly available loan dataset, trained a black-box ML model (AdaBoost) to make loan approval decisions, and employed DiCE, a popular CFE generation method, to generate candidate CFEs.

In both studies, participants were presented with multiple *CFE scenarios*. Each *CFE scenario* consisted of a rejected loan application paired with two alternative CFEs, and for each CFE scenario, participants were asked to select the recourse they would prefer to act upon. Both studies were designed to evaluate whether users (participants) prefer recourses with lower-weighted proximity, where weights are either shared across the study population or personalized for each user.

Both studies required participants to take part in two one-on-one Zoom interview sessions conducted on separate days, with each session lasting approximately one hour. This two-day format enabled us to infer individualized feature weights for each participant (on Day 1) and subsequently evaluate whether personalized-weighted proximity aligned closely with user preferences (on Day 2). Prior to beginning the pairwise comparisons, participants were given a description of CFEs and the overall loan-rejection task domain. Participants were also given the opportunity to ask clarifying questions after being shown an example *CFE scenario*. An overview of the complete workflow used across both studies is shown in Fig.2.

4.2 Pilot Study

Before running the main studies, we conducted an initial pilot user study to derive a set of global feature weights that are then used within our main studies to (i) evaluate global weighted proximity (where all users share a global set of weights); and (ii) to provide a starting point for running both our main studies (as explained in Section 4.6). To collect preference data for estimating these weights, we designed a Qualtrics survey comprising 20 loan-rejection CFE scenarios, each consisting of a rejected loan application along with two CF recourses generated using the controlled setting (described in Section 4.3). More details on how these 20 CFE scenarios were generated are provided in the appendix.

The pilot survey was conducted with 20 US-based adult participants recruited via Amazon Mechanical Turk (MTurk), each compensated with \$3 USD. Participants were first provided with a brief explanation of CFEs and the overall study task before being directed to the Qualtrics survey.

For each rejected loan scenario, participants were asked to select the recourse they would prefer to implement in order to get approved for the loan. After each selection, they were asked to provide a brief justification for their choice of recourse. After all 20 scenarios were completed, participants were asked to rate each feature on a 5-point Likert scale based on their willingness to modify it, with 1 indicating ‘very willing to modify’ and 5 indicating ‘would not modify’.

The Likert-scale responses were aggregated across participants to obtain population-level preferences for feature modification. For each feature, we computed the average willingness to modify (across all 20 participants) and normalized them to get global feature weights. Features perceived as harder to modify were assigned larger weights, whereas features perceived as easier to modify were assigned smaller weights. The resulting global weights assigned the largest weights to features pertinent to loan datasets such as ‘*education level*’, followed by ‘*employment status*’, ‘*income*’, ‘*credit score*’, and ‘*loan amount*’, in decreasing order.

4.3 Study 1: Controlled Synthetic Setting

For the first study, we sought a controlled setting in which we could precisely construct recourse pairs with known properties and minimize the effects of extraneous factors in complex real-world data. To achieve this, we designed a synthetic dataset and used an interpretable decision tree model to classify loans and generate CFEs.

Synthetic Dataset Creation. We constructed a synthetic dataset simulating binary loan application outcomes (approved vs. rejected) to maintain greater experimental control. Although several real-world credit datasets exist (Hofmann 1994; FICO 2018), they often include high-dimensional, noisy, or hard-to-understand features that can introduce confounding factors and increase the cognitive load on participants. Since our primary objective was to evaluate how lay users perceive and compare CF recourses, a simplified feature space allowed participants to focus on the core trade-offs without being overwhelmed by extraneous details.

In our synthetic dataset, each data point represents a loan application with five continuous and categorical features commonly used in lending: *income*, *credit score*, *employment status* (categorical), *education level* (categorical), and the *loan amount requested*. Feature values were independently sampled from predefined, realistic ranges that reflect plausible applicant profiles. To simulate real-world variability, we implemented multiple rule sets, each capturing a different combination of feature thresholds and decision logic for loan approval or rejection. Due to space constraints, full details of the sampling and labeling procedures are provided in the Appendix.

Training the Prediction Model. We trained a decision tree classifier on our synthetic dataset to predict loan approval outcomes. This model served as the basis for generating CFEs, which were later presented to participants.

While complex models such as neural networks might offer higher predictive accuracy, we deliberately chose decision trees for their transparency, interpretability, and rule-based structure. These characteristics made them a better fit for our study, where the goal was to evaluate how users perceive and reason about counterfactual recourses. In contrast, black-box models require the use of sophisticated, opaque CFE generation algorithms (we experiment with black-box models in study 2).

The decision tree model was trained using a standard 80/20 train-test split of the synthetic dataset and achieved 99.15% accuracy on the test set, confirming that the rule sets effectively captured the feature-label relationships. Additional details of our training procedure are provided in the Appendix.

Counterfactual Generation. Unlike most black-box ML models, decision trees allow us to trace the data point’s sequence of feature-based splits from the root to the leaf node. Let L denote a leaf node in our decision tree, and let $\phi(L) \in \{\text{accept}, \text{reject}\}$ denote the label assigned to all data points that reach leaf node L .

Our strategy for generating CFEs for our decision tree classifier (illustrated in Fig.1) leverages a key structural insight: for each rejected loan application x that reaches leaf node L (with $\phi(L) = \text{reject}$), we identify CFEs by finding paths to nearby leaf nodes L' where $\phi(L') = \text{accept}$. Each path corresponds to a set of feature changes required to shift the decision tree’s prediction from rejection to approval. For example, in Fig.1, a CFE x'_1 is constructed for input x by applying the feature modifications suggested along the path between x and x'_1 . This way, we formulate CFE generation for decision trees as a structured search over the tree, where nearby oppositely labeled leaf nodes are identified by enumerating candidate target leaves and comparing their decision paths with that of the original instance to extract the minimal set of differing decision constraints.

Using this simple rule-based algorithm, we generate CFEs for all test set instances classified as ‘rejected’ by the loan prediction model ($N=519$). For each ‘rejected’ instance x , we extract the top- K CFEs corresponding to distinct accepting leaf nodes L'_i (i.e., $\phi(L'_i) = \text{accept}$ for all $i \in 1, \dots, K$). These K CFEs are ranked in increasing order of weighted proximity to x . To ensure realism and interpretability of generated recourses, we discard CFEs that suggest implausible or counterintuitive transformations. For example, continuous features such as *income* and *credit score* are only allowed to increase in a valid recourse, reflecting feasible, forward-looking improvements in creditworthiness. For each point x , we continue generating candidate CFEs until we identify $K=15$ such realistic and semantically meaningful CFEs.

From this filtered set of K candidate CFEs for each x , we construct recourse pairs for the user study. We compare candidate CFEs and identify pairs in which one recourse has lower weighted proximity while the other has higher weighted proximity under a given weighting scheme. Each resulting *CFE scenario* consists of a rejected loan application paired with two alternative CFEs presented to partici-

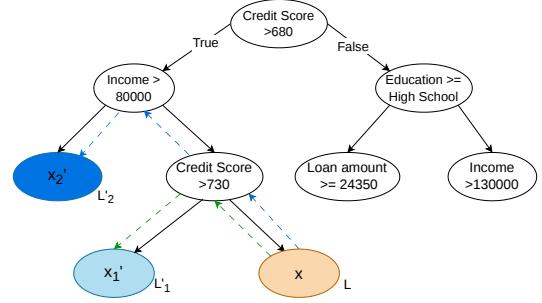


Figure 1: CFE generation with a Decision Tree Classifier. Let x denote a rejected loan application that reaches the red leaf L in the decision tree. In order to generate counterfactual recourses x'_1 and x'_2 (that reside in leaves L'_1 and L'_2 , respectively), the shortest path between leaves L and L'_1 (or L'_2) needs to be traversed. In this figure, these shortest paths are denoted using dotted green and blue paths, respectively.

pants for comparison. These recourse pairs collectively form $\mathcal{D}_1$, the dataset of *CFE scenarios* used in the first user study.

4.4 Study 2: Real-World Setting

For the second user study, our goal was to evaluate weighted proximity in a more realistic setting, without relying on synthetic datasets or transparent models. In this study, we use a publicly available loan approval dataset and train a black-box ML model to predict loan approval decisions. To generate CFEs, we use an off-the-shelf implementation of DiCE, a widely used CFE generation method (Mothilal, Sharma, and Tan 2020; Guo et al. 2026; Andringa, Baptista, and Santos 2025).

Loan Dataset. We use a publicly available loan dataset (Sharma 2023) that contains demographic and financial attributes commonly associated with credit applications. The dataset included numerical features such as applicant income, loan amount, loan term, credit score, number of dependents, and asset value, as well as categorical features like employment status and educational background. The output label is binary, indicating whether ‘loan status’ was approved or rejected.

The dataset was based on loan applications from India, so several pre-processing steps were applied before training a model to generate data points for presentation to US-based participants. Monetary values in Indian rupees were converted to US dollars, and CIBIL credit scores were rescaled from the 300-900 range to the 300-850 FICO range, which is more commonly understood by US participants. Continuous features were normalized, and categorical/binary features were one-hot encoded before being used to train the loan approval ML model.

Prediction Model. Following pre-processing, the dataset was split into 80/20 train-test sets and used to train ML models to predict loan application outcomes. To determine which model to use, several classification models were evaluated,

including logistic regression, XGBoost, random forests, gradient boosting, AdaBoost, and SVMs. Model selection was based on predictive performance on the test set, particularly the F_1 score, accuracy, and prediction stability. AdaBoost demonstrated the best overall performance ($F_1 = 0.9851$, $Acc = 98.13\%$) as compared to Gradient Boosting, which was the next best baseline ($F_1 = 0.9737$, $Acc = 96.72\%$). Therefore, AdaBoost was selected as the model for predicting loan approvals.

Counterfactual Generation. For generating CFEs, we employ DiCE (Diverse Counterfactual Explanations) (Mothilal, Sharma, and Tan 2020), a model-agnostic recourse generation framework. Given a rejected loan scenario, the DiCE framework generates a set of diverse candidate recourses using a genetic algorithm-based search procedure that identifies low-cost feature changes capable of flipping the AdaBoost prediction model’s decision from rejection to approval. The genetic search within DiCE is formulated as an optimization problem that jointly maximizes recourse validity and minimizes the proximity to the original loan instance, subject to feasibility constraints on feature modifications (e.g., the number-of-dependents feature is immutable).

For each rejected loan application in the test set, we generated a diverse pool of candidate CFEs using DiCE. Similar to the first study, we ensure that recourses generated are realistic and practical by filtering out those with counterintuitive modifications. The remaining CFEs associated with the same rejected loan application were then compared to identify recourse pairs in which one recourse had a lower weighted proximity than the other. These recourse pairs collectively form $\mathcal{D}_2$, the dataset of *CFE scenarios* used in the second user study.

4.5 Participant Recruitment

Next, participants were recruited using a combination of convenience and snowball sampling methods, including targeted recruitment emails periodically shared in student listservs at a public research university in the United States (name withheld for anonymity), posts on a public research participant recruitment platform managed by the same university in the United States, personal outreach, and social media posts. Additionally, we encouraged the first set of interested prospective participants (who replied to our recruitment messages) to share the study invitation within their personal and professional networks to increase reach and diversity. Eligibility was restricted to individuals aged 18 years or older and currently residing in the United States.

Through these efforts, a total of 43 participants were recruited for the first user study, of whom 41 completed both sessions. For the second user study, 41 participants were recruited, of whom 40 completed both sessions. The 2 participants from the first study and 1 participant from the second study who did not complete both sessions were excluded from all analyses. Participants in the first user study were compensated with a \$30 USD Amazon gift card, whereas participants in the second study received a \$25 USD Amazon gift card. *All user study and pilot study protocols were*

approved by an Institutional Review Board at the authors’ institution.

4.6 Common User Study Design

Both user studies followed the same two-phase study design, with each phase conducted on a separate day, as illustrated in Fig. 2. The only difference between the two studies was how the *CFE scenario* datasets ($\mathcal{D}_1$ from Study 1, and $\mathcal{D}_2$ from Study 2) used to construct the surveys were generated. Study 1 utilized a controlled synthetic loan dataset with decision-tree-generated CFEs to generate $\mathcal{D}_1$, whereas Study 2 used a publicly available loan dataset with CFEs generated using DiCE on an AdaBoost prediction model to generate $\mathcal{D}_2$.

Day 1. As shown in Fig. 2, on Day 1 of our user study, all participants were asked to complete a common pairwise recourse comparison task using 25 and 20 *CFE scenarios* randomly sampled from $\mathcal{D}_1$ (in Study 1) and $\mathcal{D}_2$ (in Study 2), respectively. All these Day 1 *CFE scenarios* were generated with global feature weights derived from the pilot study. For each *CFE scenario*, participants were asked to select the recourse they preferred and provide a brief explanation describing the reasoning behind their choice.

Additionally, after participants completed all pairwise comparisons across all *CFE scenarios* shown to them on Day 1, they were asked to rank the features in order of their willingness to modify them to obtain loan approval. These rankings were collected to qualitatively assess whether participants shared a similar notion of the difficulty of feature modification and whether their stated rankings were consistent with the pairwise choices made across surveys in both phases.

After this survey, each participant’s pairwise responses from Day 1 were used to infer individualized feature weights via the Bradley-Terry model as proposed by Rawal and Lakkaraju (2024). From this phase of the study, we infer participant-specific feature weights representing the perceived difficulty of modifying features in a loan application.

Day 2. Using the individualized feature weights inferred from participants’ Day 1 responses, we constructed personalized *CFE scenarios* for Day 2. For each participant, candidate CFEs associated with rejected loan applications were generated and compared to identify recourse pairs in which one CFE had a lower personalized weighted proximity score than the other. We generated surveys by randomly sampling from rejected loan applications for which at least one valid recourse pair existed for that participant.

Since personalized feature weights differed across participants, each participant received a unique set of personalized *CFE scenarios* during Day 2. Participants were presented with these surveys and asked to select their preferred recourse, but providing explanations for their selections was optional. To minimize presentation bias, we randomized the positions of the two recourses across all scenarios, ensuring that the lower-weighted scenario was not consistently placed in the same position.

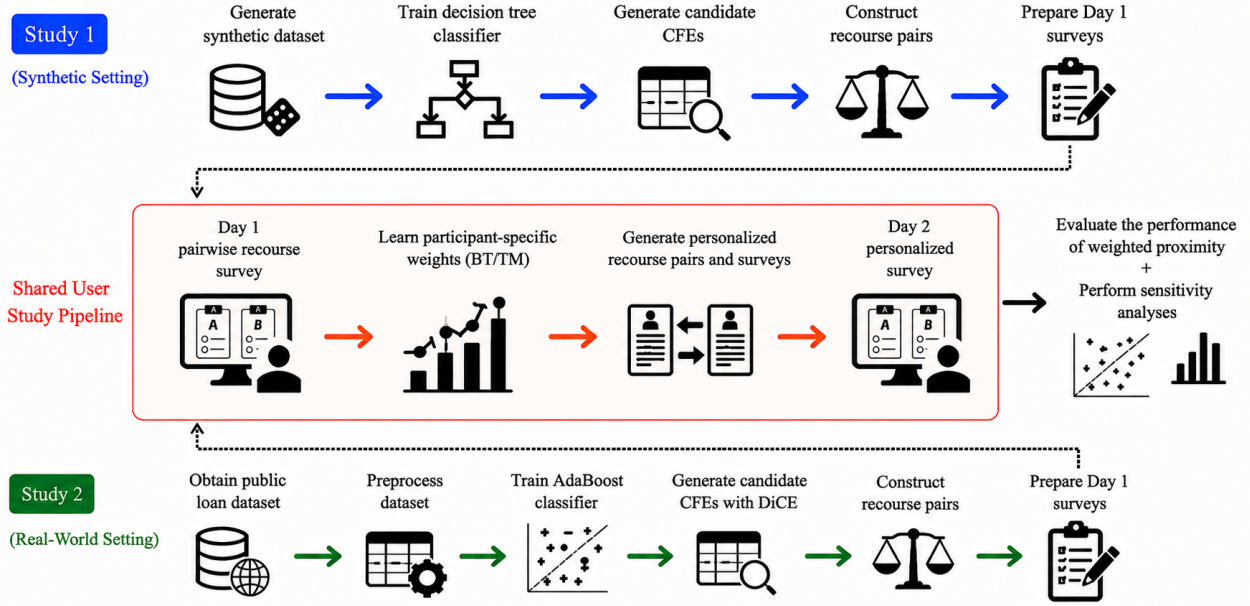


Figure 2: Workflow of our Two-Phase User Study

5 Results

Using the methodology described in the previous section, we evaluate whether weighted proximity aligns with user preferences under both global and personalized weighting schemes. We also examine whether these findings remain consistent across different experimental settings through a range of sensitivity analyses.

5.1 Global Weighted Proximity

We first evaluate whether a shared feature weighting can capture user preferences over CF recourses. On Day 1 of user study 1, participants were presented with 25 *CFE scenarios* in which weighted proximity scores were computed using global feature weights derived from the pilot study. Across all Day 1 *CFE scenarios*, participants selected the recourse with lower globally weighted proximity in only $54.05 \pm 14.11\%$ of cases (554/1025). This rate is only marginally better than chance-level selection, indicating that global feature weights derived from population-level preferences do not effectively capture individual user preferences.

User study 2 utilized a public dataset with a different set of features than those in the pilot study. Consequently, global feature weights derived from the pilot study could not be applied to *CFE scenarios* in study 2. Therefore, the global weighted proximity for study 2 was evaluated using equal weighting, assigning identical weights to all features. Across all 20 Day 1 *CFE scenarios* for study 2, participants selected the recourse with lower globally weighted proximity in only $52.38 \pm 12.40\%$ of cases (419/800), again demonstrating a near chance-level alignment between globally weighted proximity and participant preferences.

This finding is further supported by participants' responses to the feature-ranking question at the end of Day

1, in which they ranked features according to their preference for modification. The responses demonstrated substantial variability in preference rankings (across participants) and showed little to no consensus on an ordering of features. For instance, in study 1, participant P1-17 ranked their preferences as Loan Amount Requested < Income < Credit Score < Employment Type < Education, whereas P1-26 ranked them as Credit Score < Loan Amount Requested < Education < Employment Type < Income. These inconsistencies indicate that users differ significantly in how they evaluate the feasibility of feature modifications, thereby limiting the effectiveness of a single global weighting scheme.

Overall, these results demonstrate that globally weighted proximity does not consistently align with user preferences (53.21% alignment on average across both user studies) and is insufficient for effectively capturing perceptions of recourse feasibility.

5.2 Personalized Weighted Proximity

Next, we evaluate whether learning participant-specific feature weights improves alignment between weighted proximity and user preferences. In both studies, personalized feature weights were inferred for each participant using their Day 1 pairwise recourse choices and the BT model.

In user study 1, each participant was presented with personalized Day 2 surveys comprising 24 personalized *CFE scenarios* constructed using their inferred feature weights. Across all Day 2 scenarios, participants selected the recourse with lower personalized weighted proximity in $65.65 \pm 12.94\%$ of cases (646/984). While this represents an improvement over the global weighted baseline, the overall result does not indicate strong alignment between personalized feature weighting and participant preferences.

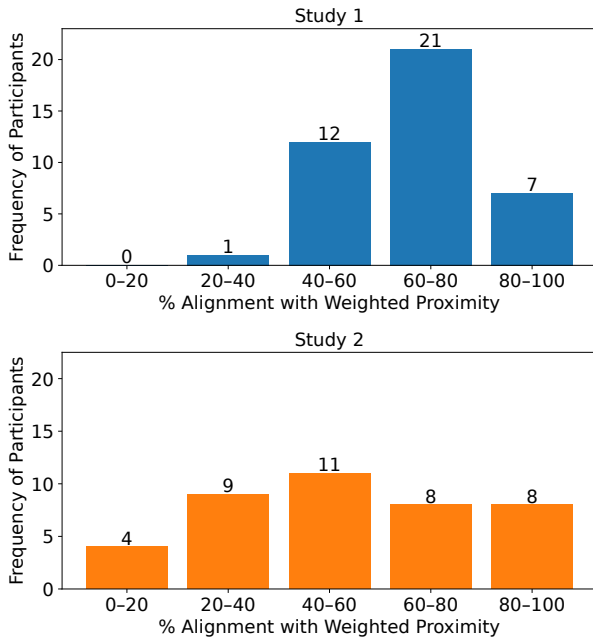


Figure 3: Participant Level Histograms of Percentage Alignment with Weighted Proximity

Since study 1 was conducted in a controlled synthetic setting, user study 2 was designed to determine whether similar findings would emerge in a more realistic setting. In Study 2, each participant received a personalized Day 2 survey comprising 15 unique *CFE scenarios*. Across all Day 2 scenarios in study 2, participants selected the recourse with lower personalized weighted proximity in only $48.17 \pm 24.34\%$ of cases. Thus, selection of the lower-weighted recourse occurred at near-chance levels (56.91% on average across both user studies), which shows a weak alignment between weighted proximity and user preferences.

Finally, Fig. 3 shows the histograms of participants' percentage match with weighted proximity in both user studies. The x-axis shows increasing percentage of alignment with weighted proximity, whereas the y-axis shows the frequency of participants. Both these figures look approximately Gaussian-shaped, which shows that most user study participants (in both studies) achieve (at best) moderate levels of alignment with weighted proximity (60-80% in Study 1 and 40-60% in Study 2), with a considerably smaller portion of participants achieving either extremely high or low alignment values.

Taken together, the findings from both studies suggest that although personalized feature weighting offers modest improvements over global weighting in controlled settings, weighted proximity does not consistently align with user preferences.

5.3 Sensitivity Analyses

The low alignment rates observed in both studies could reflect limitations of our experimental design rather than a failure of weighted proximity to capture user preferences. To

assess this possibility, we conducted three sensitivity analyses examining whether the results could be attributed to i) insufficient contrast between *CFE* pairs, ii) inadequate data for learning personalized weights, and iii) the choice of pairwise preference-learning model.

Proximity Gap Between *CFE* Pairs. One possible explanation for the observed misalignment is that the *CFE* pairs in the *CFE scenarios* presented to participants were too similar, making them practically indistinguishable. If this were the case, we would expect accuracy to improve as the weighted proximity gap between the two *CFE* pairs increases. To evaluate this, we plot the matches and mismatches from Day 2 against the absolute weighted proximity difference between the two *CFE* pairs A and B in each *CFE scenario* ($|WeightedProx(x, A) - WeightedProx(x, B)|$) across all participants. We additionally compute the Spearman correlation between the weighted proximity gap and the participant's selection of the lower-weighted recourse to assess whether there is a positive monotonic relationship between them.

Fig. 4 visualizes the relationship between the weighted proximity gap and the % alignment of weighted proximity with user preferences. The x-axis indicates the absolute weighted proximity difference between two recourse pairs A and B for a *CFE scenario*, whereas the y-axis indicates whether a participant selected the lower-weighted *CFE* (1 = match) or not (0 = mismatch). Each point on the graph represents a single Day 2 *CFE scenario*, and the dashed red line shows the best-fit trend line.

As we observe in Fig. 4, there is no meaningful relationship between the percentage alignment (y-axis) and the weighted proximity gap between the recourse pairs (x-axis). In study 1, the Spearman correlation across all day 2 *CFE scenarios* is -0.0439 ($p = 0.1691$), with an R^2 of 0.0094. The extremely small R^2 value indicates that the weighted proximity gap has almost no statistically significant effect on the match rate with weighted proximity to user preferences. Moreover, the negative correlation indicates that match accuracy decreases with increasing proximity gap. Study 2 also demonstrates a similarly weak correlation, with a Spearman correlation of 0.0261 ($p = 0.5237$) and an R^2 value of 0.0025 across all Day 2 scenarios.

The results from both studies suggest that insufficient contrast between recourse pairs is not a cause for the poor alignment between weighted proximity and user preferences.

Inadequate Pairwise Data. An argument can be made that 20-25 pairwise comparisons collected on Day 1 may be insufficient to fit individualized BT models that learn reliable feature weights. If inadequate training data were the cause of misalignment (between user responses and weighted proximity), we would expect the percentage alignment between user responses and weighted proximity to improve as the number of samples used to train the BT model increases (which, in turn, leads to more accurately learned feature weights). To investigate this possibility, we conducted a bootstrapped 5-fold cross-validation experiment using pairwise comparison data from both Day 1 and Day 2 to create a larger pool of training samples per participant.

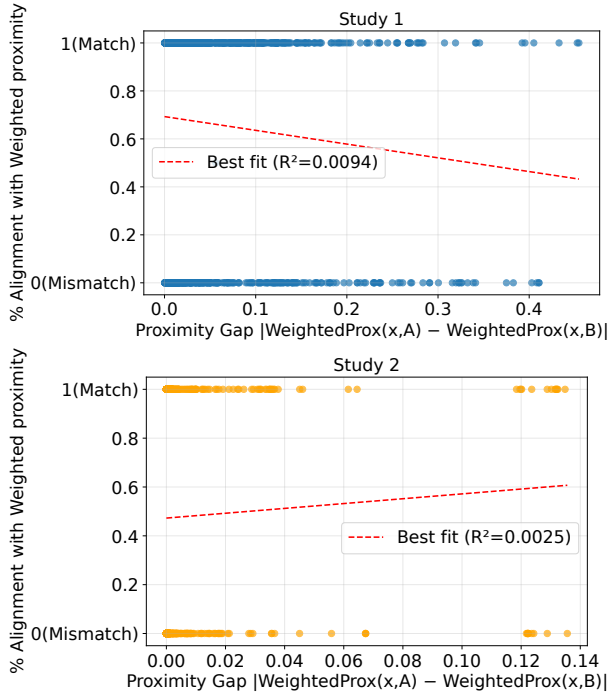


Figure 4: Variation in Percentage Alignment with Weighted Proximity vs Weighted Proximity Gap

For each participant, subsets of increasing size were randomly sampled from this pool and partitioned into five folds. In each cross-validation run, four folds were used to train the BT model, while the remaining fold was used for evaluation. This process was repeated such that each fold served as the evaluation set once. Alignment accuracy was computed for every fold and then averaged across folds for each participant and training size. This procedure was repeated for all participants to obtain the mean percentage alignment with weighted proximity at each training size.

Fig. 5 presents the resulting learning curves for the BT model across both studies. The x-axis indicates the number of pairwise comparisons used to train the feature-weight learning model, and the y-axis shows the mean *percentage alignment with weighted proximity* across participants. The solid line denotes the mean alignment (blue representing study 1 and orange representing study 2), while the shaded region represents the standard deviation across participants. For instance, in user study 1, the point at training size 28 indicates that when the BT model was trained on 28 pairwise recourse comparisons, the recourse with lower weighted proximity matched participants’ selections with an average accuracy of $70.52 \pm 13.67\%$ across all participants.

From Fig. 5, we observe that the learning curves remain relatively flat as the training size increases in both studies. These results indicate that performance improves only slightly, if at all, as additional pairwise comparisons are used to fit the BT model. In Study 1, the BT model maintains an alignment of approximately 70% across all training sizes, whereas Study 2 shows a slight decline in alignment as train-

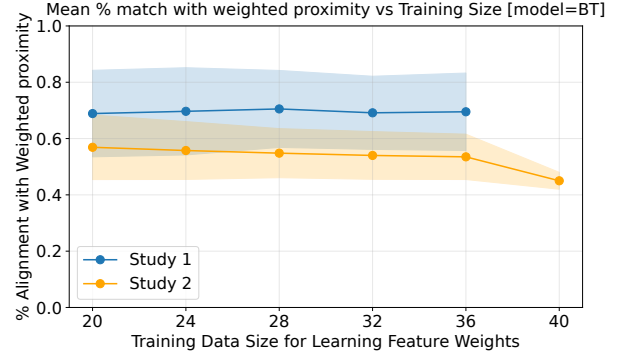


Figure 5: Result of 5-fold CV on Bradley-Terry model

ing size increases. These findings suggest that increasing the number of pairwise-comparison samples does not substantially improve the BT model’s ability to align weighted proximity with participants’ recourse preferences. Therefore, insufficient training data is unlikely to be a major cause of the weak relationship between personalized weighted proximity and user preferences.

Choice of Preference Model. Another possible explanation for misalignment is that the BT model may be poorly suited for learning personalized feature weights from pairwise-comparison data. If this were the case, we would expect predictive alignment to improve under a better alternative preference-learning framework. To address this, we replaced the BT model with the Thurstone-Mosteller (TM) model to learn feature weights as described in section 3.1. We repeated the same bootstrapped 5-fold cross-validation experiment from the previous analysis.

Fig. 6 presents the resulting learning curves for the TM model across both studies. Similar to Fig. 5, the x-axis indicates the number of pairwise training samples, while the y-axis corresponds to the percentage alignment with weighted proximity (averaged across participants). The solid line denotes the mean percentage alignment (blue representing study 1 and orange representing study 2), while the shaded region indicates the standard deviation across participants.

The TM model achieves a mean percentage alignment of $65.6 \pm 15.8\%$ across 25 training points in study 1 and $56.2 \pm 12.6\%$ across 20 training points in study 2. Although TM yields slightly higher mean percentage alignment with weighted proximity as compared to BT in study 2, the overall performance remains comparable across the two models. As shown in Fig. 6, the TM accuracy curves closely resemble those obtained using BT, with no substantial improvement in performance as additional training samples are introduced. These findings suggest that the weak relationship between personalized weighted proximity and participant preferences is unlikely to be attributed solely to the choice of preference-learning model.

6 Discussion

Weighted Proximity does not consistently align with user preferences. This work provides one of the first empirical

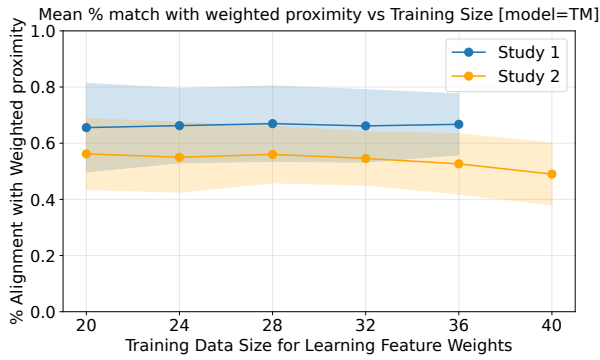


Figure 6: Result of 5-fold CV on Thurstone-Mosteller model

evaluations of weighted proximity as a proxy for user preferences in CFE generation. While weighted proximity has been discussed in prior CF literature as a method for modeling recourse feasibility, our findings suggest that both global and personalized weighting schemes exhibit weak and inconsistent alignment with how users actually evaluate and select recourses.

Across two independent user studies, participants did not consistently prefer recourses with lower weighted proximity scores. In the first study, personalized weights yielded modest improvements over global weighting in a controlled synthetic setting. However, these improvements did not generalize to the second study, which mimicked a real-world setting by utilizing a publicly available dataset and a black-box ML model. Collectively, these results indicate that weighted proximity alone may be insufficient for capturing how users reason about recourse feasibility.

Rethinking Feature-Level Cost Modeling. One possible explanation for the observed misalignment is that existing formulations of CFE objectives rely on overly simplistic assumptions that are inadequate for modeling human recourse preferences. For instance, the weighted proximity metric implicitly assumes that individuals evaluate recourses based on feature-level modification costs, where certain features are treated as harder to modify than others. Within this framework, recourse feasibility is represented as a weighted combination of feature changes.

The Need for Context Dependent Evaluation of CFEs. More broadly, our findings in this paper raise an important question about whether user-centered evaluation of CFEs can be adequately represented through a single static cost metric at all. The behavior of participants observed in our user studies demonstrated that users often evaluated CFEs holistically rather than solely through feature-level modifications (that were used to calculate a single “cost of change” metric). Although participants ranked their preferences for which features they were more or less willing to change, their selections across *CFE scenarios* frequently diverged from these stated preferences. For example, participants frequently accepted recourses involving substantial changes to features they had previously rated as difficult to modify, while rejecting recourses with minor changes to features they had considered preferable to modify. These findings

suggest that the perceived cost of modifying a feature is not fixed but instead depends on the broader context and composition of the recourse. As such, this indicates that static feature weighting may be insufficient as a proxy for user preferences. Instead, the findings of this paper suggest the need for developing evaluation criteria for CFEs that incorporate holistic and situational judgments about the recourse as a whole.

7 Limitations and Future Work

Both of our user studies were conducted exclusively in the loan approval domain, consistent with prior work in algorithmic recourse (Tominaga, Yamashita, and Kurashima 2024; Rawal and Lakkaraju 2024). Although loan approval is one of the most widely studied and high-impact settings in recourse research, future work should investigate whether similar patterns of misalignment between weighted proximity and user preferences also emerge in other recourse domains, such as hiring, criminal justice, and healthcare.

Although Study 2 utilized a publicly available dataset and a black-box model, participant evaluations were still conducted in a controlled survey environment rather than in authentic high-stakes decision-making contexts. As a result, the participant behavior observed in these studies may not reflect how individuals assess recourses when confronted with genuine financial risk, social barriers, or the long-term consequences of implementing recommended changes.

The scope of our study was based on the assumption that participant preferences could be represented using relatively stable feature-level weights. However, participant responses varied substantially across scenarios, indicating that recourse preferences may be context-dependent rather than governed by a single consistent weighting function. These findings suggest that static feature-level weighting schemes may be inadequate for modeling how users reason about recourse feasibility in practice. Future research should explore the development and evaluation of more dynamic and context-aware recourse generation models that better capture scenario-specific and holistic user reasoning.

8 Conclusion

This work addresses a significant gap within CFE research regarding the limited user-centric validation of CF generation objectives. We present one of the first empirical evaluations of weighted proximity as a metric for modeling user preferences in CF recourse generation. The results from our two studies indicate that weighted proximity exhibits weak and inconsistent alignment with human preferences under both global and personalized weighting schemes. We performed a range of sensitivity analyses whose results confirm that this poor alignment is robust across different experimental settings and modeling choices. Overall, these results underscore the need to empirically validate recourse objectives with real users and motivate future research toward more human-centered approaches to modeling recourse preferences.

Ethics Statement

Prior to any user studies, all study protocols and procedures were approved by an Institutional Review Board at the authors' Institution. Participation in our studies was voluntary, and to ensure participant well-being, they were informed that they could withdraw at any time without any consequences or loss of benefits. Before any interviews began, the study procedure and domain were explained to each participant, and verbal consent was obtained. To protect participants' privacy, all data from the interviews and surveys were securely stored on a password-protected, limited-access folder without any personally identifiable information. Moreover, only aggregated results were presented in our paper, ensuring that no results can be traced back to any individual participant.

Acknowledgments

Choudhury, Guo, and Yadav were supported in this work by the National Science Foundation under Grant No. 2426026. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- Andringa, J.; Baptista, M. L.; and Santos, B. F. 2025. Counterfactual explanations for remaining useful life estimation within a Bayesian framework. *Information Fusion*, 118: 102972.
- Assaad, S.; Zeng, S.; Tao, C.; Datta, S.; Mehta, N.; Henao, R.; Li, F.; and Carin Duke, L. 2021. Counterfactual Representation Learning with Balancing Weights. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 1972–1980. PMLR.
- Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; and Shadbolt, N. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356206.
- Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4): 324–345.
- Brughmans, D.; Leyman, P.; and Martens, D. 2023. NICE: an algorithm for nearest instance counterfactual explanations. *Data Min. Knowl. Discov.*, 38(5): 2665–2703.
- Byrne, R. M. J. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6276–6282. International Joint Conferences on Artificial Intelligence Organization.
- Carreira-Perpiñán, M. Á.; and Hada, S. S. 2021. Counterfactual Explanations for Oblique Decision Trees: Exact, Efficient Algorithms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8): 6903–6911.
- De Peuter, S.; Zhu, S.; Guo, Y.; Howes, A.; and Kaski, S. 2024. Preference Learning of Latent Decision Utilities with a Human-like Model of Preferential Choice. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 123608–123636. Curran Associates, Inc.
- Dodge, J.; Liao, Q. V.; Zhang, Y.; Bellamy, R. K. E.; and Dugan, C. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, 275–285. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362726.
- Fernández, R. R.; Martín de Diego, I.; Aceña, V.; Fernández-Isabel, A.; and Moguerza, J. M. 2020. Random forest explainability using counterfactual sets. *Information Fusion*, 63: 196–207.
- FICO. 2018. FICO (HELOC) Dataset. <https://community.fico.com/s/explainable-machinelearning-challenge?tabset-3158a=2>.
- Goodman, B.; and Flaxman, S. 2017. European Union Regulations on Algorithmic Decision Making and a “Right to Explanation”. *AI Mag.*, 38(3): 50–57.
- Guo, H.; Choudhury, F. A.; Chen, T.; and Yadav, A. 2026. Watermarking Counterfactual Explanations. In Mei, Y.; Qian, C.; Bai, Q.; Xue, B.; and Khanna, S., eds., *PRICAI 2025: Trends in Artificial Intelligence*, 51–67. Singapore: Springer Nature Singapore. ISBN 978-981-95-7081-2.
- Guo, H.; Nguyen, T. H.; and Yadav, A. 2023. CounterNet: End-to-End Training of Prediction Aware Counterfactual Explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23*, 577–589. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.
- Guo, H.; Xiong, X.; Zhang, W.; and Yadav, A. 2024. ReLax: Efficient and Scalable Recourse Explanation Benchmarking using JAX. *Journal of Open Source Software*, 9(103): 6567.
- Hofmann, H. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5NC77>.
- Huang, Y.; Li, J.; Li, M.; et al. 2023. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC Medical Research Methodology*, 23: 268.
- Karimi, A.-H.; Barthe, G.; Balle, B.; and Valera, I. 2020. Model-Agnostic Counterfactual Explanations for Sequential Decisions. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108 of *Proceedings of Machine Learning Research*. PMLR.
- Keane, M. T.; Kenny, E. M.; Delaney, E.; and Smyth, B. 2021. If Only We Had Better Counterfactual Explanations:

- Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 4466–4474.
- Liu, J.; Wu, X.; Liu, S.; and Gong, S. 2025. Model-agnostic counterfactual explanation: A feature weights-based comprehensive causal multi-objective counterfactual framework. *Expert Systems with Applications*, 266: 126063.
- Mahajan, D.; Tan, C.; and Sharma, A. 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *CoRR*, abs/1912.03277.
- Maystre, L.; and Grossglauser, M. 2011. The Thurstone-Mosteller Model from a Modern Perspective. Technical Report ADA543806, Defense Technical Information Center.
- Mc Grath, R.; Costabello, L.; Le Van, C.; Sweeney, P.; Kamiab, F.; Shen, Z.; and Lecue, F. 2018. Interpretable Credit Application Predictions With Counterfactual Explanations. In *NIPS 2018 - Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*. Montreal, Canada.
- Mersha, M.; Lam, K.; Wood, J.; AlShami, A. K.; and Kalita, J. 2024. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599: 128111.
- Mosteller, F. 1951. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1): 3–9.
- Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery.
- Rawal, K.; and Lakkaraju, H. 2020. Beyond individualized recourse: interpretable and interactive summaries of actionable recourses. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.
- Rawal, K.; and Lakkaraju, H. 2024. Learning Recourse Costs from Pairwise Feature Comparisons. arXiv:2409.13940.
- Schut, L.; Key, O.; Mc Grath, R.; Costabello, L.; Sacaleanu, B.; Corcoran, M.; and Gal, Y. 2021. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 1756–1764. PMLR.
- Sharma, A. 2023. Loan Approval Prediction Dataset. Accessed: 2025-08-08.
- Thurstone, L. L. 1927. A law of comparative judgment. *Psychological Review*, 34(4): 273–286.
- Tominaga, T.; Yamashita, N.; and Kurashima, T. 2024. Re-assessing Evaluation Functions in Algorithmic Recourse: An Empirical Study from a Human-Centered Perspective. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, 7913–7921. International Joint Conferences on Artificial Intelligence Organization.
- Upadhyay, S.; Lakkaraju, H.; and Gajos, K. Z. 2025. Counterfactual Explanations May Not Be the Best Algorithmic Recourse Approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, 446–462. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713064.
- Ustun, B.; Spangler, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, 10–19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.
- Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K.; Dickerson, J.; and Shah, C. 2024a. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *ACM Comput. Surv.*, 56(12).
- Verma, S.; Singh, G.; Mate, A.; Verma, P.; Gorantla, S.; Madhiwalla, N.; Hegde, A.; Thakkar, D.; Jain, M.; Tambe, M.; and Taneja, A. 2024b. Increasing Impact of Mobile Health Programs: SAHELI for Maternal and Child Care. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15594–15602.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2): 841–887.
- Yadav, P.; Hase, P.; and Bansal, M. 2022. Low-Cost Algorithmic Recourse for Users With Uncertain Cost Functions. arXiv:2111.01235.
- Zhang, W.; Guo, H.; Ranganathan, P.; Patel, J.; Rajasekharan, S.; Danayak, N.; Gupta, M.; and Yadav, A. 2023. A Continual Pre-training Approach to Tele-Triaging Pregnant Women in Kenya. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 14620–14627.