Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding

Peter M. VanNostrand
pvannostrand@wpi.edu
Worcester Polytechnic Institute
Worcester, USA

Dennis M. Hofmann
dmhofmann@wpi.edu
Worcester Polytechnic Institute
Worcester, USA

Lei Ma
lma5@wpi.edu
Worcester Polytechnic Institute
Worcester, USA

Elke A. Rundensteiner
rundenst@wpi.edu
Worcester Polytechnic Institute
Worcester, USA

ABSTRACT

Automated decision-making systems are increasingly deployed in domains such as hiring and credit approval where negative outcomes can have substantial ramifications for decision subjects. Thus, recent research has focused on providing explanations that help decision subjects understand the decision system and enable them to take *actionable recourse* to change their outcome. Popular *counterfactual explanation* techniques aim to achieve this by describing alterations to an instance that would transform a negative outcome to a positive one. Unfortunately, little user evaluation has been performed to assess which of the many counterfactual approaches best achieve this goal. In this work, we conduct a crowd-sourced between-subjects user study (N = 252) to examine the effects of counterfactual explanation type and presentation on lay decision subjects' understandings of automated decision systems. We find that the region-based counterfactual type significantly increases objective understanding, subjective understanding, and response confidence as compared to the point-based type. We also find that counterfactual presentation significantly effects response time and moderates the effect of counterfactual type for response confidence, but not understanding. A qualitative analysis reveals how decision subjects interact with different explanation configurations and highlights unmet needs for explanation justification. Our results provide valuable insights and recommendations for the development of counterfactual explanation techniques towards achieving practical actionable recourse and empowering lay users to seek justice and opportunity in automated decision workflows.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Computing methodologies** → **Artificial intelligence**.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658997

KEYWORDS

Explainable AI, User Studies, Algorithmic Transparency, Human-Computer Interaction.

ACM Reference Format:
Peter M. VanNostrand, Dennis M. Hofmann, Lei Ma, and Elke A. Rundensteiner. 2024. Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3630106.3658997

1 INTRODUCTION

As machine learning systems have grown more capable, they have rapidly been deployed to automate decision-making tasks in consequential domains such as finance [35, 41], recruitment [16, 32], healthcare [37], and policing [13, 23] where negative decisions can have substantial impacts on decision subjects' lives. Motivated by this alarming trend, explainable AI (XAI) techniques have been developed to provide decision subjects with an understanding of how a decision is made, and thus the possibility of taking recourse [52]. Requirements for explanation of automated decisions are also increasingly being codified into law [1, 41, 43].

Of particular interest have been so-called *counterfactual explanations* as they are believed to meet legal requirements for lay user appropriate explanation [61]. These explanations provide decision subjects with actionable recourse for undesired negative outcomes (e.g., the denial of a loan) by describing alterations to the features of their instance that would lead to a positive outcome (e.g., suggesting a loan applicant increase their income to some amount to obtain an approval) and are best suited to decisions on tabular data [26]. There has been a flurry of activity in this area resulting in different notions of counterfactual explanation [54]. Approaches vary both in *counterfactual type*, such as point-based counterfactuals (e.g., *income* = \$1,000) [12, 42, 50, 56] and region-based counterfactuals (e.g, \$1,000 < *income* < \$1,500) [17, 20, 60] as well *counterfactual presentation* with different styles for explaining the same content, ranging from simple numeric capture [58], to textual description [53], and visual depictions [18].

Unfortunately, while counterfactual explanation is frequently supported by drawing parallels to human notions of reasoning [2,

36], surveys of the field have found that in practice the design of XAI techniques is driven by machine learning experts with little grounding in psychology and without a thorough investigation of real users' needs [39, 40]. As a result, comparative evaluation of counterfactual approaches is largely limited to computational metrics such as counterfactual proximity and distributional faithfulness, with multiple competing metrics often intending to measure the same notion [19, 28]. While interesting, these computational metrics do not capture the understanding and needs of lay users whose knowledge and priorities have been shown to differ significantly from those of machine learning experts [22].

Existing explanation user studies have not yet bridged this gap, with a recent survey finding that only a handful of studies consider counterfactuals [49]. Of these, most compare counterfactuals to other forms of explanation (e.g., feature importance). Such works have found that counterfactual explanations can increase metrics of understanding [62] and improve perceptions of fairness [51, 64] and justice [55] as compared to non-counterfactual methods. Despite these promising results, existing studies are restricted to singular configurations of counterfactual, typically point-based counterfactuals presented as text (Sec. 2.2). Thus, a large unmet need remains for user studies to examine to what degree decision subjects understand and may be able to use different types of counterfactual explanations for actionable recourse and to determine what presentation styles are most easily understood. Motivated by this need, we address the following three research questions in this work:

- **RQ1:** What effect do counterfactual explanation type and counterfactual explanation presentation have on lay decision subjects' understanding of automated decision-making systems?
- **RQ2:** What effect do counterfactual explanation type and counterfactual explanation presentation have on lay decision subjects' confidence in their understanding of automated decision-making systems?
- **RQ3:** Do lay decision subjects' subjective understanding and confidence predict their objective task performance?

We develop six unique counterfactual explanation configurations varying across two key factors: counterfactual type (*point-based* vs *region-based*) and counterfactual presentation style (*numeric*, *natural language*, and *visual*). We examine a population of lay users acting as simulated decision subjects for a loan approval scenario, and propose a methodology (Sec. 4) to evaluate their understanding by presenting a series of loan decisions alongside counterfactual explanations. Following this procedure, we perform a crowd-sourced 2x3 between-subjects user study with $N = 252$ participants. We measure participants' *subjective understanding* via agreement statements and *objective understanding* via accuracy across twelve task questions in three recourse-related areas. We also record participants' *response confidence* and *response time* for task questions, and solicit user experiences via open response.

Using this information, we perform a quantitative analysis for eight internally preregistered hypotheses (Sec. 5.1) followed by an exploratory statistical analysis (Sec. 5.2) and a qualitative analysis of open responses (Sec. 6). Our analysis finds significant effects of counterfactual type, with region-based counterfactuals leading to significantly higher objective understanding, subjective understanding, and response confidence than point-based counterfactuals. Further, we find counterfactual presentation style does not significantly effect user understanding in this context, but does significantly effect response time and moderates the effect of counterfactual type on participants' response confidence. We also show that users' subjective understanding and response confidence are significant predictors of their objective understanding. Based on our results, we provide recommendations (Sec. 7) for XAI practitioners to focus on the development and deployment of practical region-based counterfactual explanation techniques. We also encourage HCI researchers to continue exploration of presentation methods to determine how best to maximize understanding.

2 BACKGROUND AND RELATED WORK

Counterfactual explanations are a form of post-hoc local explanation: post-hoc in that they are generated after a machine learning model is trained, and local in that they are specific to a particular instance [19]. Counterfactual explanations are answers to a counterfactual question typically formulated as *"Why P rather than Q"* where P is some factually observed outcome, typically undesired (e.g., loan denial), and Q is some hypothetical counterfactual outcome desired by the user (e.g., loan approval) [36]. Counterfactual explanation is thought to follow human notions of reasoning with literature from psychology finding people typically value why one event happened rather than another [2, 36]. Counterfactuals have also been argued to meet emerging regulatory requirements for lay user appropriate explanation of consequential decisions [61] though the scope and enforcement of such regulation remains an emerging area of law [3].

2.1 Counterfactual Explanation Methods

Point-Based Counterfactuals. Counterfactual explanation has largely been explored in the context of single counterfactual points often called counterfactual examples. That is, given an instance $x \in \mathbb{R}^n$ predicted as class P, point-based counterfactual explanation methods seek to find some hypothetical point $x' \in \mathbb{R}^n$ which would be predicted as class Q. Numerous methods to generate counterfactual points have been explored, including via algorithmic search [14, 45, 50, 56], linear programming [10, 25, 42], and gradient access [12, 31, 38]. These methods guide the generation of x' by metrics such as *similarity*, often the L1 or L2 norm between x and x'; *sparsity*, the number of features differing between x and x'; and *validity*, the reliability of x' obtaining the desired outcome [19]. Counterfactual validity is of particular importance to the human context because methods that fail to guarantee validity may waste a user's time and effort if they make changes to meet x' and re-receive an unwanted outcome. This also risks liability for the owner of the decision system under relevant regulatory frameworks [1, 41, 43].
Region-Based Counterfactuals. Recently, methods have emerged for creating counterfactual explanations that cover a portion of the values in the feature space larger than just a single point. These region-based approaches provide greater flexibility to users by offering additional context and information on the rationale of the decision system. This aligns with observations that point-based

counterfactuals may place unrealistic requirements on users to precisely set the value of each feature in spite of normally expected feature variability (e.g., requiring a loan applicant obtain a very specific bank account balance) [27]. These methods utilize similar abstractions to capture portions of the feature space, such as describing a set of disjoint rules across one or more features, e.g., one rule option restricts ($1000 < *income* < $1500) while another option restricts ($500 < *rent* < $900), or via a continuous hyperbox e.g., ($1000 < *income* < $1500 AND $500 < *rent* < $900) [15, 17, 20, 60]. These region-based methods differ in two key ways: 1) whether all points that fall into the region are guaranteed to be validly counterfactual, and 2) the computational complexity required for creating the region explanation. Specifically, LORE [20] and LEWIS [17] fail the validity guarantee, and RFOCSE's [15] core approach has been shown to be intractable for reasonably sized ensembles [60]. This led RFOCSE to adopt a faster heuristic-variant which lacks a validity guarantee. In Sec. 4 we use the recent work FACET [60] to generate counterfactuals as it provides a strong guarantee of counterfactual validity in efficient time.

2.2 Explanation User Studies

Non-Counterfactual Studies. Existing user studies on explanation largely focus on feature importance techniques [49] such as those generated by LIME [34] and SHAP [48]. Cheng *et al.* [6] examined local feature importance explanations presented as stacked bars compared to a lack of explanation for user understanding in a college admissions scenario. They found the presence of explanations led to higher objective understanding measured by three model simulation tasks, but not higher subjective understanding measured via post-task questionnaire. Poursabzi-Sangdeh *et al.* [44] examined global feature importance explanations and found their presence led users to more accurately predict the model's behavior, but made them worse at detecting its mistakes.

Point-Based Studies. Some works compare the effects of understanding from counterfactuals, but consider only point-based counterfactual types. For example, Wang *et al.* [62] found that both feature importance and point-based counterfactual explanations presented as structured text improve objective understanding with similar effect size for recidivism prediction, but observed only an increase in subjective understanding for counterfactuals in a forest cover scenario. Bove *et al.* [4] use a loan application scenario and found that multiple counterfactual points shown via a card-style UI improved objective and subjective understanding compared to single counterfactual points. This suggests that additional counterfactual information may improve understanding. Warren *et al.* [63] compare textual counterfactual points using categorical and continuous features for automated drunk driving assessment and found objective understanding to be higher with categorical features.

Studies have also evaluated the effect of point-based counterfactuals with other metrics. Kuhl *et al.* [29] compared closest counterfactual points to plausible counterfactual points using a numeric presentation and found closest counterfactuals lead to faster learning of an abstract game [30]. Schoeffer *et al.* [51] and Yurrita *et al.* [64] examined the effect of combined counterfactual and feature importance explanations presented as text to other forms of explanation. They found the combined explanations led to the highest user perceptions of fairness. Binns *et al.* [55] examine perceptions of justice with differing forms of explanation presented via text and found that counterfactuals led to higher perceptions of justice, but that these effects were outweighed by scenario effects. While the above works each adopt a different explanation presentation style, none directly examine this factor.

Presentation Studies. Works which directly evaluate explanation presentation are more rare. Van Berkel *et al.* [59] compared textual data summary explanations to visual data scatterplots for loan applications and recidivism prediction, and found the text cases yielded higher perceived fairness. Szymanski *et al.* [55] compared textual, visual, and hybrid presentations of multiple forms of explanation for article reading time estimation. They found that hybrid forms of explanation led to the highest objective understanding. No works to date have adequately studied the effect of presentation for counterfactual explanations.

3 RESEARCH HYPOTHESES

To explore our primary research questions (Sec. 1), we pose the following hypotheses.

3.1 Hypotheses for RQ1: User Understanding

Wang *et al.* [62] show that point-based counterfactuals can improve user understanding of an automated decision system and Bove *et al.* [4] find that the information from multiple counterfactual points can improve both objective and subjective understanding compared to a single point. Because region-based explanations contain a super-set of the information provided by point-based counterfactuals, and indeed enclose many counterfactual points, we expect that region-based counterfactuals would thus increase both objective understanding (H_{1a}) and subjective understanding (H_{1c}) compared to point-based counterfactuals. Further, Szymanski *et al.* [55] find some evidence that explanation presentation may impact objective understanding. Therefore, we anticipate that presentation may moderate the effect of counterfactual type on objective understanding (H_{1b}). Finally, van Berkel *et al.* [59] find that subjective perceptions of fairness differ between explanation presentations. Therefore, we predict that subjective understanding of the automated decision-making system may also differ between presentation styles (H_{1d}).

- **Hypothesis 1a (H_{1a})**: Region-based counterfactual explanations improve objective user understanding as compared to point-based counterfactual explanations.
- **Hypothesis 1b (H_{1b})**: The effect of explanation type on objective understanding is moderated by explanation presentation.
- **Hypothesis 1c (H_{1c})**: Region-based counterfactual explanations improve subjective understanding as compared to point-based counterfactual explanations.
- **Hypothesis 1d (H_{1d})**: Users' subjective understanding differs based on explanation presentation.

3.2 Hypotheses for RQ2: User Confidence

When an instance (e.g., a loan application) does not exactly match a point-based counterfactual, little definitive information is available to the user about what the outcome of that instance would be. This may cause users to be uncertain or otherwise feel that they are

guessing. In contrast, region-based counterfactuals provide greater information on the decision-making system which may be used to assess the likelihood of alternate outcomes, e.g., by simply determining if the instance falls within the region (Sec. 2.1). Therefore, we hypothesize that region-based counterfactuals may lead users to have greater confidence (i.e., belief/certainty in their understanding) than point-based counterfactuals (H_{2a}). Additionally, if users perceive one explanation presentation to be more complex or difficult to parse than another, they may be less confident using that presentation (H_{2b}). We can measure this confidence by providing users a spectrum of responses and recording how often they choose the more extreme responses.

- **Hypothesis 2a (H_{2a})**: Users are more confident in their responses with region-based counterfactual explanations than with point-based counterfactual explanations.
- **Hypothesis 2b (H_{2b})**: The effect of explanation type on users' response confidence is moderated by explanation presentation.

3.3 Hypotheses for RQ3: Calibrated Understanding

For users to effectively use counterfactual explanations for actionable recourse, they must be able to accurately assess their level of understanding of the underlying decision-making system. This is critical as users who are "confidently wrong" may expend significant effort enacting a set of changes which will not result in the desired counterfactual outcome. Thus, users' subjective understanding and confidence are ideally *calibrated* with their true objective understanding. While Cheng *et al.* [6] and Poursabzi-Sangdeh *et al.* [44] observe some divergence in objective and subjective understandings for feature importance explanations, evaluations by Bove *et al.* [4] and Warren *et al.* [63] of different uses of counterfactuals showed increases in both. This may indicate that when working with counterfactuals users have a fairly accurate self-assessment of their understanding. Therefore, we expect that subjective understanding (H_{3a}) and response confidence (H_{3b}) are positively associated with objective understanding for both types of counterfactual explanation.

- **Hypothesis 3a (H_{3a})**: Users' subjective understanding is positively associated with objective understanding
- **Hypothesis 3b (H_{3b})**: Users' response confidence is positively associated with objective understanding

4 METHODOLOGY

To test our hypotheses about the effects of counterfactual explanation type and presentation on lay user understanding of automated decision systems (Sec. 3), we conducted a randomized human subjects experiment ($N = 252$) on Prolific [46] using explanations for mock loan approval decisions (Sec. 4.1). Six configurations of counterfactual explanation were considered varying in counterfactual type (region vs point) and counterfactual presentation (numeric vs natural language vs visual) for a complete 2x3 between-subjects design. Communication of loan decisions and explanations was operationalized through an explanation user interface (Fig. 1) customized for each configuration (Sec. 4.2). User understanding was assessed through both quantitative measures (Sec. 4.3) and thematic analysis of open response (Sec. 6). We detail our survey procedure and participant recruitment strategy in Sec. 4.4. The design of the explanation user interfaces and the exact wording of evaluation questions was refined through a series of internal user groups and a small crowd-sourced pilot study of twenty participants.

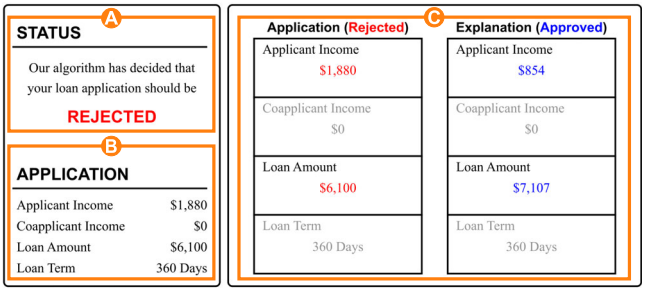


Figure 1: Sample counterfactual explanation user interface

4.1 Experiment Scenario

Following prior research [4], we adopt a loan application scenario for our experiment due to the consequential nature of such decisions, familiarity of lay users with the task, and legal requirements for explanations of automated decisions in this domain [1, 41, 43]. To generate counterfactual explanations for the experiment, we trained a machine learning (random forest) classifier on a random 80% sample of a Kaggle loans dataset [24]. This model acts as an automated decision system which predicts *Loan Approval* (binary REJECT/APPROVE) from *Applicant Income*, *Coapplicant Income*, *Loan Amount*, and *Loan Term*. We then used the model to classify the remaining 20% of applications and explained those which were Rejected (119) with the state-of-the-art technique FACET [60]. This produced a region-based and point-based counterfactual explanation for each application decision. To avoid biasing participants towards certain feature-values, we examined each explanation and selected 12 distinct instances for the study that were diverse in terms of the altered features and counterfactual values. Each instance was used exactly once.

4.2 Explanation Interface Prototypes

For each of the six counterfactual explanation configurations, we develop an *explanation user interface* to display the loan decision and associated explanation to the user. These interfaces share a common templated layout and sidebar as depicted in Fig. 1. *Area A* reminds the user of the loan scenario and rejection decision, *Area B* displays the feature-values of the given loan application, and *Area C* contains the counterfactual explanation of the decision for that application. Each of the six configurations shown in Fig. 2 combine one counterfactual type and presentation and are plugged into *Area C*. Each participant is randomly assigned to a single explanation configuration to create six experimental groups. We develop configurations for each factor as follows.

Explanation Type. We explore two types of counterfactual explanation: point-based and region-based counterfactuals.

- ***Point-Based Counterfactuals.*** For a given instance, a point-based counterfactual explanation is a set of feature-values, one per feature, such that if the instance is transformed to exactly match those values, the automated decision-making system will produce the desired counterfactual outcome.

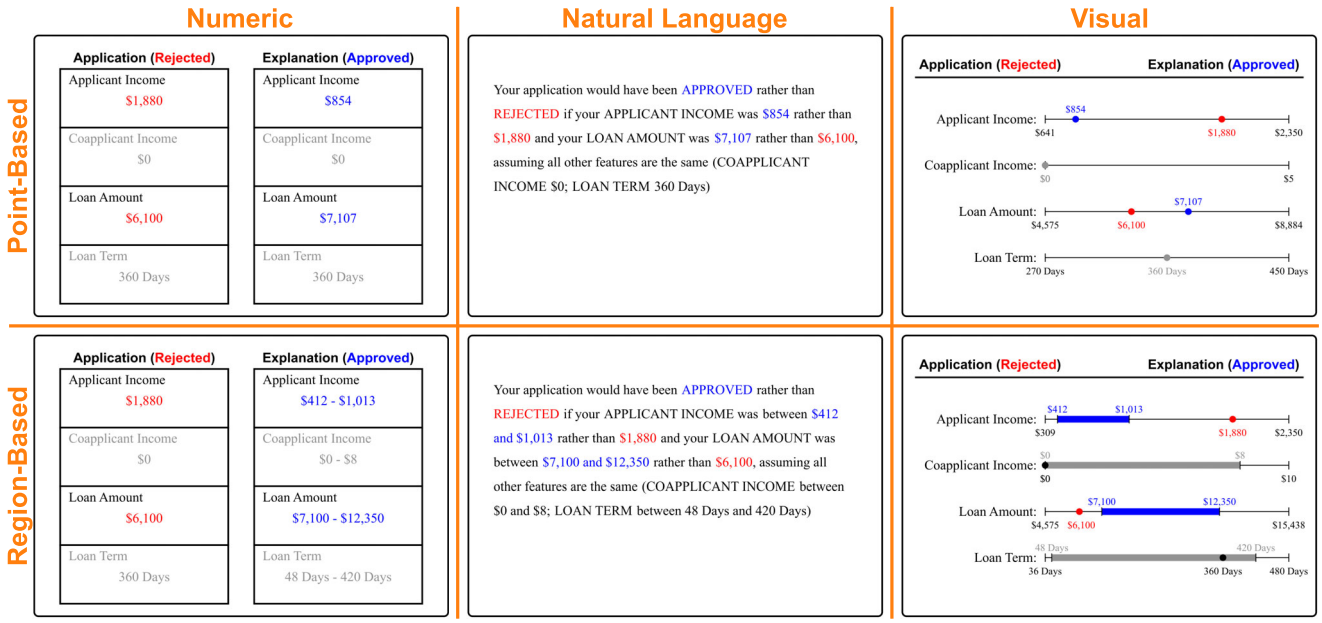


Figure 2: Explanations for one instance using the six studied counterfactual explanation configurations

- ***Region-Based Counterfactuals.*** For a given instance, a region-based counterfactual explanation is a continuous bounded range along each feature such that any point that falls within the prescribed range for every feature is guaranteed to obtain the desired outcome from the automated decision-making system.

Explanation Presentation. We explore three counterfactual explanation presentations: numeric, natural language, and visual as they represent three distinct communication modalities prevalent in existing works (Sec. 2). For each counterfactual type, all three styles contain the same counterfactual information and we adopt a set of simple and consistent design principles to mitigate confounding effects. To aid readability, features with proposed alterations have their current factual values displayed in red and the newly proposed counterfactual values shown in blue. Features requiring no alteration are shown in grey. All presentations provide feature information in alphabetical order by feature name.

- ***Numeric.*** Following existing research [29], we organize the feature-values into a tabular arrangement (Fig. 2 Left) for a structured representation and display a side-by-side comparison between the observed factual values and the counterfactual values proposed by the explanation. In the point-based case, this is simply a single value, e.g., *ApplicantIncome* : \$854, while in the region-based case, this is presented as a range, e.g., *ApplicantIncome* : \$412 – \$1,013.
- ***Natural Language.*** Existing works often present counterfactuals via text [53, 55, 64]. To examine this case, we develop a templated natural language presentation (Fig. 2 Center). Each statement begins with a description of the decision outcome and then lists the altered features indicating the prescribed counterfactual values contrasted to the factual values with a *rather than* clause. The counterfactual values for regions are provided using the word *between* to encode the range. Finally, a parenthetical listing indicates the non-altered features and their values.
- ***Visual.*** A limited set of works provide explanation content via visualization [55, 59]. To investigate this, we develop a simple visual presentation which uses number line plots (Fig. 2 Right). Here, each feature is given its own number line displaying both the factual and counterfactual values for that feature. Point-based counterfactuals are depicted using points, while region-based counterfactuals are represented by shaded bars along the line.

4.3 Evaluation Metrics

Objective Understanding (H_{1a}, H_{1b}, H_{3a}, H_{3b}). Due to the complex nature of human processing, many metrics measuring user understanding exist. Following previous HCI studies of explanations [4, 6, 7, 44, 55, 62, 63], we adopt the definition that a user "understands" a decision system if they can identify what attributes cause the system's actions and can predict how changes in the situation can lead to alternative outcomes. Following this philosophy, we adapted the evaluation questions from [4, 7] to target actionable recourse. Specifically, we designed three types of task questions to assess participants' objective understanding of the decision system in three critical recourse-related areas: *Feature Alteration*, *Instance Prediction*, and *Feature Sensitivity*.

- ***Question Type 1: Feature Alteration.*** As counterfactual explanations provide understanding of the decision system through proposed feature alterations, it is critical for actionable recourse that users can accurately identify the alterations prescribed by the explanation. Correctly interpreting this information allows a user to determine the significant features and threshold values in the local space of the explained instance. To measure this ability, we presented participants with an explanation for a not-before-seen instance and asked them via a multiple-choice question to identify *which change is most likely to get this application approved?* from among three potential choices (Appx. C.1). To account for preexisting assumptions of model behavior (e.g., participants assuming that a higher income is always more likely to be approved), we selected a mix of explanations with intuitive and counterintuitive alterations. Participants were also directed

to not rely on their preexisting assumptions and incentivized via bonus payments to answer correctly.

- ***Question Type 2: Instance Prediction.*** Another component of actionable recourse is assessing whether or not an instance will achieve the desired decision outcome. This is critical as it reflects a user's understanding of the underlying decision system's behavior and enables them to determine whether or not they have sufficiently altered their instance to match the provided counterfactual explanation. We measured this ability directly by presenting participants with an explanation for a rejected instance alongside a new instance of unknown outcome. We then asked the participants to predict the system's decision for the new instance on a 4-point forced-choice scale from *very likely to be rejected* to *very likely to be accepted.*
- ***Question Type 3: Feature Sensitivity.*** Once a user can identify the alterations a counterfactual explanation suggests and determine if their instance will achieve the desired outcome, they must "freeze" their instance (i.e., prevent significant deviation from their altered feature-values) until the decision system processes the new instance. For example, a loan applicant may abstain from large transactions to "freeze" their savings balance while they reapply for the loan. To achieve this, the user must understand what features are *sensitive.* That is, they must be able to identify features of their instance that would result in an undesired outcome if allowed to deviate by a small amount. This is equivalent to identifying which feature(s) of their instance are closest to a decision boundary. To measure this understanding, we presented participants with an explanation for a rejected instance alongside a new instance which we told them was accepted. We asked them to choose *Which attribute of your new application, if changed by a small amount, is most likely to result in a rejection?*

We created 12 objective understanding questions, 4 of each type. Response options were generated by altering the instance to meet the explanation, then adjusting one or more feature-values in each option as described in Appx. C.1. Responses were scored compared to ground truths, assigning one point to each correct answer to create an *objective understanding* score ranging 0-12.

Subjective Understanding (H_{1c}, H_{1d}, H_{3a}). In addition to measuring a user's true understanding of the model, we also measure their self-reported understanding. This is important to determine if some forms of explanations lead to a false sense of understanding which could cause users to expend effort in fruitless attempts to achieve their desired outcome. To measure subjective understanding, we adapted the questions from [4] and asked participants to indicate their agreement with five statements (Appx. C.2) on 6-point bipolar forced choice Likert-style agreement scales. We converted each response to a value 0-5 and summed the 5 questions to create a *subjective understanding* score ranging from 0-25.

Response Confidence (H_{2a}, H_{2b}, H_{3b}). For the Instance Prediction questions (Objective Question Type 2), we asked the participants to predict the outcome for an instance on a 4-point scale from *very likely to be rejected* to *very likely to be accepted.* To measure how confident users were in their responses, we computed the number of times each participant chose a *very likely* option over a *somewhat likely* option. This yields a score ranging from 0-4 as we asked four such questions.

Response Time. To approximate the difficulty of processing different explanation configurations, we measured how much time each participant spent answering each objective understanding question and computed the total response time.

Satisfaction and Trust. We asked participants to rate their satisfaction with the provided explanations on a 5-point Likert-style scale from *not satisfied* to *highly satisfied* and indicate their agreement with the statement *I trust the decisions made by the algorithm* on a 6-point forced choice scale from *strongly disagree* to *strongly agree* (Appx. C.3).

4.4 Survey Procedure and Recruitment

The experiment consisted of an online survey administered via Qualtrics [47] with the following five major steps.

1) **Presurvey.** After collecting consent, we asked participants questions relating to their demographics and individual factors. One question was manipulated to act as an attention check. We then randomly assigned each participant to one of the six counterfactual explanation configurations shown in Fig. 2.
2) **Introduction.** We gave participants a description of the loan application scenario and familiarized them with the features used. We also asked a set of simple recall questions to ensure they read the materials (Appx. C.5).
3) **Training.** We presented participants with an example explanation (e.g., Fig. 1) for the configuration corresponding to their group. We then used a short series of descriptions and questions to train the participants to locate the factual instance's values, identify altered features, and understand the criteria for acceptance provided by the explanation (Appx. C.6).
4) **Task Evaluation.** To measure objective understanding and response confidence we asked the participants to answer the twelve task questions from our three recourse-related question areas (Appx. C.1)
5) **Post Survey.** We concluded the survey by having participants complete the Likert-style agreement questions for subjective understanding, satisfaction, and trust, with one question manipulated to act as an attention check (Appx. C.2).

Based on a power analysis of the tests (Sec. 5.1) for our main hypotheses, we aimed to collect data from at least 247 participants. We thus recruited 264 participants from Prolific [46] in September 2023. All participants were adults ($\geq$ 18 years old), first-language English speakers, located in the United States. Recruitment was limited to Prolific members who had completed 100+ tasks with a 95% or higher approval rate. Prolific's "gender-balanced" recruitment feature was applied. Each participant was paid a $4 base amount and up to an additional $2 based on their accuracy in answering questions. Participants were informed of the bonus potential with the amount scaling in $0.50 increments to incentivize participants to make legitimate responses. Participants were not told which/how many questions they got correct to avoid biasing their responses. The average completion time was 20.32 minutes and average was compensation $5.07 for an average wage of $14.98/hr. For analysis, we excluded ten participants who failed at least one of the two attention checks and two participants who failed more than two

Hypotheses for Main Effects ($p < 6.25 \times 10^{-3}$ significant)		$F \uparrow$	p-value$\downarrow$	$\eta_p^2 \uparrow$
H_{1a}	Regions increase objective understanding	217.34	$<2 \times 10^{-16}$	0.4694
H_{1b}	Presentation moderates objective understanding	4.04	3.13×10^{-2}	0.0278
H_{1c}	Regions increase subjective understanding	60.91	1.71×10^{-13}	0.1984
H_{1d}	Presentation effects subjective understanding	2.10	1.25×10^{-1}	0.0168
H_{2a}	Regions increase response confidence	42.41	4.14×10^{-10}	0.1474
H_{2b}	Presentation moderates response confidence	6.18	2.41×10^{-3}	0.0478
H_{3a}	Subjective understanding predicts objective understanding	-	1.39×10^{-10}	-
H_{3b}	Response confidence predicts objective understanding	-	1.30×10^{-4}	-
Additional Related Observations				
O_1	Presentation effects objective understanding	0.29	0.7458	0.0024
O_2	Presentation moderates subjective understanding	3.53	0.0309	0.0279
O_3	Presentation effects response confidence	0.17	0.8479	0.0014

Table 1: Results of statistical tests for main hypotheses and related observations

of the simple recall questions from Procedure Steps 2-3. This left $N = 252$ participants whose demographics are shown in Appx. B.2.

5 QUANTITATIVE ANALYSIS AND RESULTS

5.1 Hypothesis Tests

Here, we perform statistical tests to evaluate our 8 main hypotheses (Sec. 3). To conservatively control the family-wise error rate to below 0.05, we apply Bonferroni correction yielding $\alpha = 0.05/8 = 0.00625$. Thus, p-values from the main analysis below are only considered significant if $p < 6.25\times10^{-3}$. For hypotheses analyzed with ANOVA tests, we report the partial eta squared (η_p^2) effect size in addition to p-value and F statistic and use Cohen's [9] rules for interpretation. The significance of all tests is shown in Tab. 1 and the main metric scores for each of the six counterfactual explanation configurations are shown in Fig. 3. We also perform Bayesian ANOVA for some tests and report Bayes Factors for these cases using Lee and Wagenmaker's [33] rules.

RQ1: Objective Understanding. Our first confirmatory analysis is a multi-way ANOVA test with counterfactual explanation type (region-based vs point-based) and presentation (numeric vs natural language vs visual) as factors predicting participants' *objective understanding* (Sec. 4.3). Here we find a large effect of explanation type (H_{1a}) with significantly higher understanding for region-based explanations ($\mu = 9.24 \pm 0.22$) than point-based explanations ($\mu = 5.46 \pm 0.14$, score ranges 0-12). We find some evidence suggestive of a small moderating effect of explanation presentation on the effect of explanation type (H_{1b}), but a Bayesian ANOVA reveals anecdotal evidence in favor of no moderating effect ($BF_{01} = 2.16$). Thus, we cannot reject the null hypothesis for this case. We also did not observe an effect of explanation presentation alone on objective understanding (O_1), with a Bayesian ANOVA revealing strong evidence for the null hypothesis that it has no effect ($BF_{01} = 22.05$).

RQ1: Subjective Understanding. A second multi-way ANOVA uses explanation type and presentation as factors predicting *subjective understanding* (Sec. 4.3). We again find a large effect of explanation type with a significantly higher mean score for region-based explanations ($\mu = 18.33 \pm 0.42$) than point-based explanations ($\mu = 13.49 \pm 0.47$, score ranges 0-25). Similarly to objective understanding, we did not find evidence for an effect of explanation presentation on subjective understanding (H_{1d}). A Bayesian ANOVA reveals moderate evidence in favor of the null hypothesis that presentation does not have an effect ($BF_{01} = 5.28$). We also observed evidence suggesting a small moderating effect of explanation presentation on explanation type for subjective understanding, (O_2), but we did not register a hypothesis for this case.

RQ2: Response Confidence. A third multi-way ANOVA uses explanation type and presentation as factors predicting *response confidence* (Sec. 4.3). Here, we find a large significant effect of explanation type on response confidence (H_{2a}) favoring region-based explanations ($\mu = 2.17\pm0.08$) over point-based explanations ($\mu = 1.31\pm0.11$, score ranges 0-4). We also find a small significant moderating effect of explanation presentation on the effect of explanation type for response confidence (H_{2b}). We did not observe evidence indicating an effect of explanation presentation alone (O_3).

RQ3: Calibrated Understanding. We performed a *multiple linear regression analysis* to test the association of participants' response confidence and subjective understanding with their objective understanding ($R^2 = 0.24$, $p = 4.82 \times 10^{-16}$, $F = 40.77$). Our results show that both subjective understanding ($H_{3a}, \beta = 0.1929$) and response confidence ($H_{3b}; \beta = 0.5400$) are both significantly positively associated with objective understanding – with response confidence being the stronger predictor.

Summary of Hypothesis Findings. We found sufficient evidence to reject the null hypothesis for 6 of our 8 tests. This includes large effects of explanation type on objective understanding, subjective understanding, and response confidence (H_{1a}, H_{1c}, H_{2a}); a moderating effect of explanation presentation on the effect of explanation type for response confidence (H_{2b}); and that subjective understanding and response confidence are both significant predictors of objective understanding (H_{3a}, H_{3b}). Finally, we found some evidence that explanation presentation may moderate the effect of explanation type on subjective understanding (H_{1b}), but could not reject the null in this case. We found no evidence that presentation alone effects subjective understanding (H_{1d}).

5.2 Exploratory Analysis

Here, we provide additional findings related to our hypothesis tests (Sec. 5.1) and examine secondary factors that may effect understanding of counterfactual explanations. We also consider additional metrics for explanation utility, as shown in Fig. 4.

Expansion on Moderation Effects. For hypothesis H_{2b}, we found a small significant moderating effect of explanation presentation on the effect of type for response confidence (Sec. 5.1). To examine this further, we performed a follow-up Tukey test. This reveals that region-based counterfactuals have significantly higher response

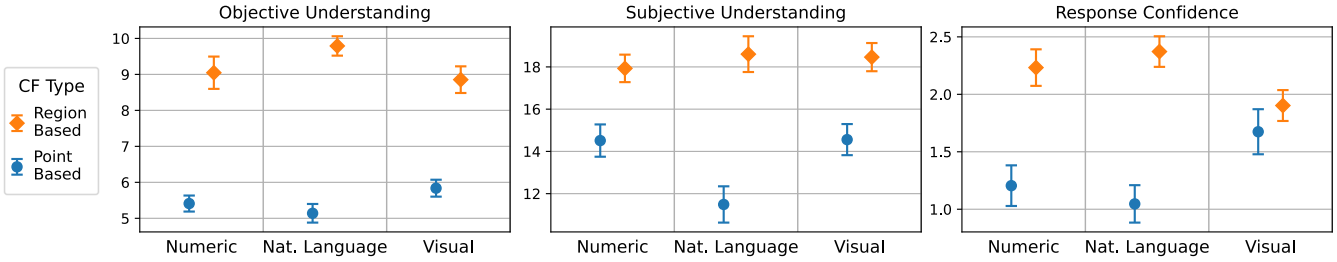


Figure 3: Mean and standard error of main metrics for the six configurations of counterfactual explanation type and presentation

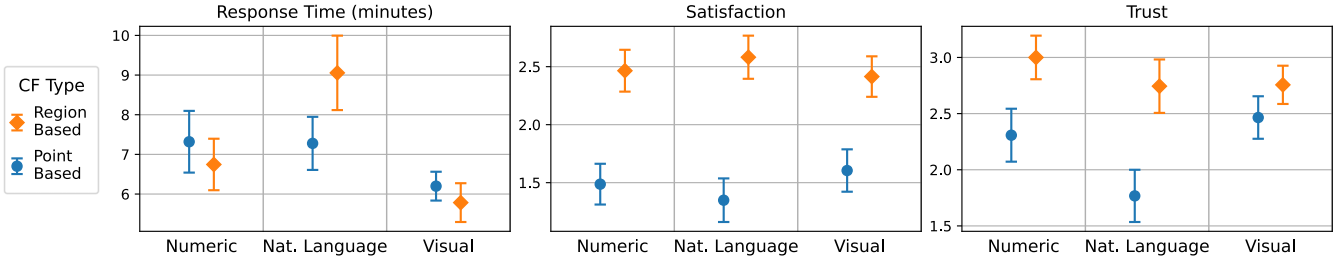


Figure 4: Mean and standard error of additional metrics for the six configurations of counterfactual explanation

confidence than point-based counterfactuals for the natural language ($p < 2 \times 10^{-7}$) and numeric ($p = 2.07 \times 10^{-4}$) presentations, but not for the visual case ($p = 0.92$). This can be seen in Fig. 3. Similarly, in O_2 we found that explanation style may moderate the effects of type for subjective understanding. A Tukey test for this metric finds a significant difference between subjective understandings from visual and natural language presentations in point-based counterfactuals ($p = 0.0475$), but not in region-based counterfactuals ($p = 0.99$).

Task Understanding. To dig into users' understanding of individual tasks, we disaggregated the objective understanding score (Sec. 4.3) into the three task areas: *feature alteration*, *instance prediction*, and *feature sensitivity*. We then repeated the multi-way ANOVA test from Sec. 5.1 for each. We find that explanation type remains significant for all three tasks (alteration $p < 2 \times 10^{-16}, \eta_p^2 = 0.33$; prediction $p = 1 \times 10^{-15}, \eta_p^2 = 0.23$; sensitivity $p < 2 \times 10^{-16}, \eta_p^2 = 0.33$). See Appx. B.1 for task scores.

Response Time. A multi-way ANOVA predicting *response time* (Sec. 4.3) with explanation type and presentation finds presentation has a small yet significant effect ($p = 5.89 \times 10^{-3}, F = 5.24, \eta_p^2 = 0.0409$) while type has no effect ($p = 0.59$). A Tukey test reveals visual presentations resulted in significantly faster response times than natural language presentations ($p = 3.94 \times 10^{-3}$) by several minutes (visual $\mu = 6.00 \pm 0.30$; language $\mu = 8.17 \pm 0.58$). Response times for numeric styles ($\mu = 7.02 \pm 0.50$) could not be significantly distinguished. Separate linear regressions found response time is not associated with subjective ($p = 0.23$) or objective understanding ($p = 0.39$), or response confidence ($p = 0.86$).

Satisfaction and Trust. We ran two multi-way ANOVAs using explanation type and presentation to predict user-reported satisfaction and trust. These revealed a large significant effect of explanation type on satisfaction ($p < 8 \times 10^{-11}, F = 42.21, \eta_p^2 = 0.1582$) and a small significant effect of type on trust ($p < 1.74 \times 10^{-4}, F = 12.54, \eta_p^2 = 0.0553$). No effects were found for explanation presentation alone on satisfaction ($p = 0.50$) or trust ($p = 0.27$).

6 QUALITATIVE ANALYSIS

Here, we analyze participant responses to a series of open response questions using reflexive thematic analysis under a constructionist framework [5, 8]. This allows us to identify latent themes in the data that provide insight into how users may conceptualize explanations and how those concepts effect their utilization. We asked four open response questions (Appx. A), one about the clarity of explanation, and three about how participants completed the task evaluation. The responses total 1,008 text passages and we analyzed responses to the "how" and "clarity" questions separately. Responses were coded using QualCoder [11] over several iterative steps to refine initial codings into the identified subthemes. All quoted data extracts Q.i are available in Appx. A, Tab. 2.

6.1 Region-Based Counterfactuals Encourage Reliable Range Checking

When participants described how they answered the recourse-driven task questions, a notion of assessing the "fitness" of an instance to an explanation was common. This fell into three subthemes: a) *ambiguous distance*, b) *range checking*, and c) *wiggle room*. Participants using point-based explanations often relied on a notion of *ambiguous distance* (45/125) by assuming that instances which were *"close"* (Q.2), *"similar to"* (Q.3), or *"nearest"* (Q.1) in distance to the counterfactual point better fit the explanation and were thus more likely to receive the desired loan approval outcome. This latent assumption is neither guaranteed nor holds for many cases, e.g., a nearby point may be across the decision boundary and get rejected while a far point is approved. It's also unclear how close is *"close enough"* (Q.6), and this likely varies widely between users.

In contrast, many fewer (14/127) participants using region-based explanations made determinations based on distance. Instead, they frequently performed *range checking* (59/127) where they evaluate if the application's values are *"within"* (Q.7) or *"fit into"* (Q.10) the explanation's range on that feature. Most participants looked for matches on all features, but those that didn't tended to map the

number of ranges met to the likelihood of approval (Q.8). This underlying belief is mathematically true, but in practice only points that satisfy all ranges are guaranteed the desired outcome. Similarly, in the sensitivity task participants using region-based explanations regularly used and even named (Q.11, Q.12) a notion of *wiggle room* (42/127). This maps neatly to the goal of identifying features near decision boundaries, with participants referencing the *"limit"* (Q.13) *"maximum"* (Q.14), or *"borderline"* (Q.15) of the region's range. Combined, these patterns demonstrate the value of region-based counterfactuals to provide criteria that are easily understood by real users and which help resolve issues of ambiguous distance underlying the use of point-based counterfactuals.

6.2 Users Seek Justification For Counterfactuals That Don't Match Their Assumptions

When asked to describe the clarity of the given explanations, a pattern of informational understanding combined with justification seeking was common. This pattern is present through parallel subthemes of a) *action clarity*, and b) *assumptions of reasoning*. For example, one participant wrote *"I think the explanation of why the loan was rejected is clear, but WHY those criteria are valid does not make sense"* (Q.20). Indeed, participants generally found the information given by the explanations to be easy to understand, with many (74/252, excluding single word yes/no responses) giving wholly positive descriptions of the explanation UI. This included strong indications that they understood the suggested alterations (Q.18, Q.19) and responses highlighting that specific characteristics such as the use of color were helpful (Q.16, Q.17).

Despite this, many participants (69/252) expressed confusion over why the specific counterfactual values were selected. In particular, participants pointed at *"counterintuitive"* (Q.21) suggestions (e.g., increasing the loan amount to obtain approval) to not *"make sense"* (Q.22) or to be *"illogical"* (Q.22) and sought further explanations of why these changes were suggested (Q.24). This reveals that users have strong underlying assumptions about how an automated decision-system works – namely that such systems do or should closely follow human reasoning. In practice many counterintuitive counterfactual changes are possible as machine learning systems are not constrained to follow such notions. Further, an explanation that seems counterintuitive or unreasonable at first glance may have a rationally grounded underpinning – e.g., a microlending service rejecting applicants whose incomes are too high, or a bank rejecting a loan amount that is too small to be profitable. Without such a justification, participants question explanations that deviate from their assumptions, with a few even raising concerns that the underlying decision system may be *"predatory"* (Q.26) or *"sketchy"* (Q.25). This may have substantial impacts on perceptions of fairness and trustworthiness.

7 DISCUSSION AND RECOMMENDATIONS

7.1 Effects of Explanation Type: Clear Wins for Region-Based Counterfactuals

Results from our hypothesis tests (Sec. 5.1) find that region-based counterfactual explanations lead to significantly higher objective understanding (H_{1a}), subjective understanding (H_{1c}), and response confidence (H_{2a}) than point-based counterfactuals among our population of lay users. Further, our exploratory analysis confirms that objective understanding remains higher for region-based counterfactuals across all three recourse-related task areas (Sec 5.2). This indicates that using region-based counterfactuals, lay users are better able to identify the required counterfactual alterations, assess the fitness of an instance with respect to those changes, and discern which features of an instance are nearest to deviating from the proposed alterations. These strong increases may be due to the more reliable process of range checking that we observe in our qualitative analysis (Sec. 6.1). Our hypothesis tests also find that participants accurately identified their increased ability to perform these tasks; with both subjective understanding and response confidence being positively associated with objective understanding (H_{3a}, H_{3b}). Finally, evidence from our exploratory analysis also finds corresponding increases in reported satisfaction and trust in the automated decision system among participants with region-based counterfactuals. In combination, these results indicate that region-based counterfactuals are well suited for use by lay users and may hold significant promise for practical actionable recourse.

To leverage these findings, machine learning experts should consider a) focusing on developing region-based explanation approaches similar to those from emerging methods [17, 20, 60] for a wider variety of model types; and b) investigating efficient methods for embedding region-based counterfactuals as part of standard practice when creating systems for high-stakes automated decision-making.

7.2 Effects of Explanation Presentation: A Call for Additional Examination

The effects of counterfactual explanation presentation are less clear than the effects of counterfactual type. Considering the results of our hypothesis tests (Sec. 5.1), we did not find significant evidence of explanation presentation interacting with explanation type on objective understanding (H_{1b}), nor evidence of explanation presentation alone having an effect on subjective understanding (H_{1d}). Similarly, exploratory analyses revealed no evidence for an effect of explanation presentation on users' reported satisfaction or trust (Sec 5.2). Implications for the lack of these effects are mixed. On one hand, observing that three different explanation presentations achieve the same level of user understanding may indicate that lay users are capable of digesting counterfactual explanation information through a variety of modalities. On the other hand, the lack of observed effects does not give a clear indication of best practices for HCI designs to maximize user understanding. This may be due in part to the intentional similarity in design of the numeric, natural language, and visual presentations we examine. To minimize confounding variables all three presentations contain identical information, apply the same color coding schema, and are presented in context of the same explanation interface. While this leads to consistent interfaces designs, we cannot ensure they are optimal and therefore an examination of more diverse presentations may reveal more significant differences in understanding.

Despite the lack of evidence for effects of presentation on user understanding, our hypothesis tests do find that explanation presentation has a small significant moderating effect on explanation type for response confidence (H_2b). Further, in our exploratory analysis,

we observe a large significant effect of presentation on response time, with participants using natural language presentations taking on average more than two minutes longer than those using visual presentations. This may be due to the need to scan the natural language presentation multiple times to locate the required explanation information, with the visual presentation allowing participants to locate the same information much more quickly. These results suggest that while explanation presentation may not improve user understanding, different presentation styles may increase or decrease the amount of effort required to reach that understanding. This may be relevant for domains with low-motivation users who may decline to expend the required effort and where rapid explanation interpretation is pertinent for decision-making.

To fully understand the effects of different explanations presentations we suggest HCI researchers examine a broader array of design options for explanation interfaces both within counterfactual explanation, and among explanation systems more broadly. While existing human studies are valuable in demonstrating the potential benefits and pitfalls of explanation, it's critical that we go beyond explanation content alone and examine how presentation may help (or harm) user experiences in explanation workflows.

8 LIMITATIONS AND FUTURE WORK

Our work has the following limitations that may be addressed through future work. First, as with many studies our findings are context specific. We examine the effects of counterfactual explanation type and presentation on lay user understanding for actionable recourse of loan application decisions. While our insights in this context are substantial, future work is needed to evaluate how these findings generalize to other domain scenarios with different factors and stakes and to consider a wider variety of presentation styles. Evaluation of different user populations should also be considered, including non-lay user groups as appropriate for the target domain. Second, our work focuses primarily on evaluating user understanding. In Sec. 5.2, we find explanation type significantly affects user trust and satisfaction as measured by single Likert-style questions. However, as these factors are important for the practical use of explanations, a more in-depth evaluation would be valuable. Other metrics such as perceptions of fairness and justice are also relevant and worth investigating. Third and lastly, our study is tailored specifically towards actionable recourse for negative decision outcomes. Such recourse relies on fundamental assumptions about the mutability of features and users' abilities to enact the proposed alterations. The mere presence of counterfactual explanations does not guarantee these to be true. Thus, this work should not be used to justify the automation of consequential decisions without careful consideration of the negative impacts on users. As automated decision-making systems increasingly determine the shape of our society, a great deal of technical and legal work remains needed to ensure that automated decision-making systems are used ethically and that their decisions can be reliably audited and fairly contested.

9 CONCLUSION

In this work, we bridge the gap between XAI methods development and user perspectives by examining how lay users experience explanations for actionable recourse of automated decisions. In particular, we perform a between-subjects user study to evaluate the effects of counterfactual explanation type and presentation on lay user understanding in a loan application scenario. Our analysis finds that region-based counterfactuals result in significantly higher objective understanding, subjective understanding, and response confidence compared to point-based counterfactuals. We also find that region-based counterfactuals lead to significantly higher user satisfaction and trust. Based on these results, we recommend machine learning experts focus on the development of these region-based counterfactual techniques and include such explanations as part of practical automated decision-making systems. Additionally, we find that explanation presentation can significantly moderate some of the above effects of explanation type, and that natural language presentations greatly increases response times compared to visual ones. Given the recent explosion of automated decision-making systems and the corresponding increase in regulatory scrutiny, our findings point to an unmnet need for HCI and fairness researchers to study how best to serve users with effective explanation information to enable diverse user populations to effectively utilize explanations across a variety of high-stakes domains.

ACKNOWLEDGMENTS

This research was supported in part by NSF under grants IIS-1910880, CSSI-2103832, CNS-1852498, NRT-HDR-2021871 and the U.S. Department of Education under grant P200A180088. Thanks also to the members of the DAISY research group.

RESEARCH ETHICS AND SOCIAL IMPACT

Ethical Considerations. We considered and addressed the following ethical factors when designing our study.

Scenario Selection. As actionable recourse is especially critical for consequential decisions, we were interested in studying the effects of counterfactual explanation for a realistic, relatively high stakes scenario. However, we did not want to use a scenario which might cause our participants undue stress. We chose not to work with the available COMPAS recidivism dataset as a carceral scenario may raise traumatic experiences for participants who have a history with the criminal legal system. Similarly, we considered using undergraduate applicant data from our institution, but avoided doing so as this may be a point of stress for some participants and could potentially involve partial disclosure of real student data, even if anonymized. More broadly, we felt that using such scenarios may inadvertently endorse or normalize the use of automated decision-making systems in these domains, where such uses remain controversial. We chose the loan application scenario because 1) the dataset is publicly available on Kaggle; 2) the scenario is reasonably but not overly consequential; and 3) automated systems are largely accepted for making such decisions. Additionally, the financial field has a comparatively long history of protective regulation to structure the use of allowable decision-making processes. We also specifically excluded dataset features related to demographic and personal background as these are not practically actionable and could raise issues of bias if used by the decision-making system.

Participant Rights and Privacy. As the field of data collection has been known to exploit crowd-sourced labor, we took the following steps to protect our participants. First, all participants were required

to complete a consent form to ensure they understood the study expectations. This included the study goal, risks/benefits, compensation, expected duration, the right to exit at any time, and contact information for our Institutional Review Board (IRB) office. Second, we used internal testing and a small pilot study to determine the average completion time and adjusted compensation to meet the minimum wage in our jurisdiction ($15/hr). Third, to respect participant's privacy we recruited respondents pseudonymously via Prolific and replaced Prolific IDs with randomized Participant IDs before analysis. All researchers also underwent CITI Program training for responsible data handling. Finally, to avoid pressuring participants, all demographic questions were optional and collected flexibly (e.g., age in range brackets, gender as open response, and multiple selectable race options including a custom option). Demographic details were collected to characterize our sample population and contextualize our results, but were not used as predictive factors for analysis. The above process and all survey content was reviewed and approved by our IRB.

Researcher Positionality. The researchers conducting this work come from a largely American background with research experience in computational solutions to human-centric data problems, and robust access to educational and technological resources. These factors inevitably influence the design of our study and the analysis of our results. Thus, our recommendations may not be equally applicable or appropriate for the use of explanations and automated decision-making systems in populations with different cultural norms or language use, and in populations where access to education and technology may be more limited.

Adverse Impacts. The findings of this work should not be seen to in any way endorse or justify the use of automated decision-making systems for high-stakes tasks. Indeed, the proliferation of machine learning systems in critical decision-making has and will continue to shape society and profoundly affect individuals lives, particularly as companies often fail to take seriously even the most basic duties of care for how such systems impact the people they touch. The mere addition of explanations like those studied in this work does not mitigate these effects and explanation should not be used to create a misplaced sense of trust or otherwise misrepresent the decision-making process. With or without explanations, automated decision systems can be used to reinforce historical patterns of marginalization, automate unjust systems of power, and foreclose opportunities for meaningful change. We therefore encourage governments, community members, and labor organizations to use explanations as only one of many tools for deeply examining such systems, and to take action when needed to ensure that if decision-making is to be automated, it is done in a way that protects the rights of decision subjects, and leads to fair and just outcomes.

REFERENCES

[1] Equal Credit Opportunities Act. 1974. Public Law, 15 C.F.R § 1691, Regulation B 12 C.F.R. § 1002.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[3] Sebastião Barros Vale and Gabriela Zanfir-Fortuna. 2022. *Automated decision-making under the gdpr: Practical cases from courts and data protection authorities.* Technical Report. Future of Privacy Forum.

[4] Clara Bove, Marie-Jeanne Lesot, Charles Albert Tijus, and Marcin Detyniecki. 2023. Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: An Explanation User Interface Proposition and User Study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 188–203. https://doi.org/10.1145/3581641.3584082

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[6] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[7] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[8] V Clarke and V Braun. 2019. *Guidelines for reviewers and editors evaluating thematic analysis manuscripts.* Technical Report. University of Auckland.

[9] J Cohen. 1988. *Statistical power analysis for the behavioral sciences* (2 ed.). Routledge, Oxfordshire, United Kingdom. https://doi.org/10.4324/9780203771587

[10] Zhicheng Cui, Wenlin Chen, Yujie He, and Yixin Chen. 2015. Optimal Action Extraction for Random Forests and Boosted Trees. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 179–188. https://doi.org/10.1145/2783258.2783281

[11] Colin Curtain. 2023. *QualCoder.* ccbogel. https://github.com/ccbogel/QualCoder/releases/tag/3.4

[12] Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. 2019. Model Agnostic Contrastive Explanations for Structured Data. *CoRR* abs/1906.00117 (2019), 12 pages. arXiv:1906.00117 http://arxiv.org/abs/1906.00117

[13] Will Heaven Douglas. 2020. Predictive policing algorithms are racist. MIT Technology Review. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

[14] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. 2022. Robust Counterfactual Explanations for Tree-Based Ensembles. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, Baltimore, MD, USA, 5742–5756. https://proceedings.mlr.press/v162/dutta22a.html

[15] Rubén R. Fernández, Isaac Martín de Diego, Víctor Aceña, Alberto Fernández-Isabel, and Javier M. Moguerza. 2020. Random forest explainability using counterfactual sets. *Information Fusion* 63 (2020), 196–207. https://doi.org/10.1016/j.inffus.2020.07.001

[16] Joseph B Fuller, Manjari Raman, Eva Sage-Gavin, and Kristen Hines. 2021. *Hidden workers: Untapped talent.* Technical Report. Harvard Business School Project on Managing the Future of Work and Accenture.

[17] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD '21)*. Association for Computing Machinery, New York, NY, USA, 577–590. https://doi.org/10.1145/3448016.3458455

[18] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: Visual Counterfactual Explanations for Machine Learning Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 531–535. https://doi.org/10.1145/3377325.3377536

[19] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 36 (2022), 1–55. https://doi.org/10.1007/s10618-022-00831-6

[20] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *CoRR* abs/1805.10820 (2018), 10 pages. arXiv:1805.10820 http://arxiv.org/abs/1805.10820

[21] Jennifer L Hughes, Abigail A Camden, Tenzin Yangchen, et al. 2016. Rethinking and updating demographic questions: Guidance to improve descriptions of research samples. *Psi Chi Journal of Psychological Research* 21, 3 (2016), 138–151.

[22] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[23] Angwin Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[24] Kaggle. 2008. Loan Predication. https://www.kaggle.com/datasets/ninzaami/loan-predication,.

[25] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. 2020. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, Yokohama, Kanto, Japan, 2855–2862. https://doi.org/10.24963/ijcai.2020/395 Main track.

[26] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.* 55, 5, Article 95 (dec 2022), 29 pages. https://doi.org/10.1145/3527848

[27] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3442188.3445899

[28] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. arXiv:2103.01035 [cs.LG]

[29] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2125–2137. https://doi.org/10.1145/3531146.3534630

[30] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2023. Let's go to the Alien Zoo: Introducing an experimental framework to study usability of counterfactual explanations for machine learning. *Frontiers in Computer Science* 5 (2023), 20.

[31] Thai Le, Suhang Wang, and Dongwon Lee. 2020. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 238–248. https://doi.org/10.1145/3394486.3403066

[32] Colin Lecher. 2019. How Amazon automatically tracks and fires warehouse workers for 'productivity'. The Verge. https://www.theverge.com/2019/4/25/18516004/amazon-warehouse-fulfillment-centers-productivity-firing-terminations

[33] Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press, Cambridge, England. https://doi.org/10.1017/CBO9781139087759

[34] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach, CA, USA. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[35] Enmanual Martinez, Lauren Kirchner, and The Markup. 2021. The secret bias hidden in mortgage-approval algorithms. Associated Press. https://apnews.com/article/lifestyle-technology-business-race-and-ethnicity-mortgages-2d3d40d5751f933a88c1e17063657586

[36] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[37] Beth Mole. 2023. UnitedHealth uses AI model with 90% error rate to deny care, lawsuit alleges. Ars Technica. https://arstechnica.com/health/2023/11/ai-with-90-error-rate-forces-elderly-out-of-rehab-nursing-homes-suit-claims/

[38] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. https://doi.org/10.1145/3351095.3372850

[39] Shane T. Mueller, Elizabeth S. Veinott, Robert R. Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J. Clancey. 2021. Principles of Explanation in Human-AI Systems. *CoRR* abs/2102.04972 (2021), 10. arXiv:2102.04972 https://arxiv.org/abs/2102.04972

[40] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55, 13s, Article 295 (jul 2023), 42 pages. https://doi.org/10.1145/3583558

[41] CFPB Newsroom. 2023. CFPB Issues Guidance on Credit Denials by Lenders Using Artificial Intelligence. Consumer Financial Protection Bureau. https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/

[42] Axel Parmentier and Thibaut Vidal. 2021. Optimal Counterfactual Explanations in Tree Ensembles. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Virtual, 8422–8431. https://proceedings.mlr.press/v139/parmentier21a.html

[43] Article 29 Data Protection Working Party. 2016. Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679. https://ec.europa.eu/newsroom/article29/items/612053

[44] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[45] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 344–350. https://doi.org/10.1145/3375627.3375850

[46] Prolific. 2023. Prolific crowsourcing platform. https://www.prolific.com. Accessed: 2023-12-04.

[47] Qualtrics. 2023. Qualtrics Experience Management. https://www.qualtrics.com. Accessed: 2023-12-04.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[49] Y. Rong, T. Leemann, T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci. 5555. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 36, 01 (nov 5555), 1–20. Issue 33. https://doi.org/10.1109/TPAMI.2023.3331846

[50] Maximilian Schleich, Zixuan Geng, Yihong Zhang, and Dan Suciu. 2021. GeCo: Quality Counterfactual Explanations in Real Time. *Proc. VLDB Endow.* 14, 9 (oct 2021), 1681–1693. https://doi.org/10.14778/3461535.3461555

[51] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1616–1628. https://doi.org/10.1145/3531146.3533218

[52] Gesina Schwalbe and Bettina Finzel. 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* 37 (2023), 1–59. Issue 1.

[53] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* 5, 8 (2023), 873–883. https://doi.org/10.1038/s42256-023-00692-8

[54] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

[55] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[56] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 465–474. https://doi.org/10.1145/3097983.3098039

[57] Meng-Jung Tsai, Ching-Yeh Wang, and Po-Fen Hsu. 2019. Developing the Computer Programming Self-Efficacy Scale for Computer Literacy Education. *Journal of Educational Computing Research* 56, 8 (2019), 1345–1360. https://doi.org/10.1177/0735633117746747 arXiv:https://doi.org/10.1177/0735633117746747

[58] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. https://doi.org/10.1145/3287560.3287566

[59] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 245, 13 pages. https://doi.org/10.1145/

3411764.3445365

[60] Peter M. VanNostrand, Huayi Zhang, Dennis M. Hofmann, and Elke A. Rundensteiner. 2023. FACET: Robust Counterfactual Explanation Analytics. *Proc. ACM Manag. Data* 1, 4, Article 242 (dec 2023), 27 pages. https://doi.org/10.1145/3626729

[61] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard journal of law & technology* 31, 2 (2017), 841–.

[62] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[63] Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. 2023. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 171–187. https://doi.org/10.1145/3581641.3584090

[64] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 134, 21 pages. https://doi.org/10.1145/3544548.3581161

A OPEN RESPONSE AND SELECTED QUOTES

Selected quotes from the qualitative analysis (Sec. 6). Quotes are referenced to anonymous participant identifier and the question it responds to. "How" questions are labeled by which of the three recourse-related areas (Sec. 4.3) they describe and are responses to the question *How did you use the explanation tool to answer these questions?* This was asked three times during Task Evaluation (Procedure Step 4), once after each block of questions from the corresponding area. The "clarity" question *How did you use the explanation tool to answer these questions?* was asked after Training (Procedure Step 3, Sec. 4.4).

Q.i	Quote	Participant
Q.1	*chose the nearest values of the changes suggested listed in the multiple choice*	P172-Alter
Q.2	*I looked to see if the numbers were close*	P44-Pred
Q.3	*tried to pick the closest numbers that correlate to the numbers on the approved side*	P72-Alter
Q.4	*comparing the number and looking for number that were close or the same*	P34-Pred
Q.5	*If the numbers for the applicant were similar to the approved numbers from the algorithm, I felt the chances of being approved would be higher*	P91-Pred
Q.6	*I judged whether the stats were "close enough" to the algorithms preferences*	P135-Pred
Q.7	*If all 4 criteria dont fit within the approved junction parameters, I would say it would be rejected*	P10-Pred
Q.8	*If they fit into every blue category they were very likely to be approved. If they fit into most of them they were fairly likely, etc*	P174-Pred
Q.9	*I checked for each change and looked to see if it was within the range*	P170-Alter
Q.10	*I checked to see if the changes fit into the amounts the algorithim listed*	P237-Alter
Q.11	*By deducing which category had the least amount of wiggle room to be changed*	P14-Sense
Q.12	*I looked at the parameters of each section and chose the area that had the least amount of "wiggle room"*	P14-Sense
Q.13	*I looked at the ranges and saw which was closest to the limit*	P48-Sense
Q.14	*If the new numbers were close to being at the minimum or maximum of the approvals*	P188-Sense
Q.15	*If the new applicant stats were borderline to being in the rejected zone I chose those*	P67-Sense
Q.16	*easy to see that red portions are rejected and how things need to change in order to become blue and approved*	P51-Clear
Q.17	*The blue color makes it easy to understand the necessary changes that will get your application approved.*	P25-Clear
Q.18	*Yes, it is very clear about what needs to be changed for me to get approved for the loan*	P65-Clear
Q.19	*It's easy to understand why I was rejected and what I would need to do in order to be accepted*	P188-Clear
Q.20	*the explanation of why the loan was rejected is clear, but WHY those criteria are valid does not make sense*	P141-Clear
Q.21	*It is easy to read, but what the tool is suggesting that you do seems counterintuitive*	P169-Clear
Q.22	*It does not make sense to require less income for a larger loan than the applicant applied for*	P183-Clear
Q.23	*It was confusing because it wasn't intuitive. Why would a loan agency want to give you more money while you make less? It seems illogical*	P27-Clear
Q.24	*I would like to see more explanation about why the algorithm thinks that certain number will allow for approval*	P231-Clear
Q.25	*No. Seems sketchy, though*	P231-Clear
Q.26	*it almost seems like it is incentivizing a predatory nature where it prefers applicants that earn less money and take higher loan amounts to get more profit at the expense of putting them in debt*	P176-Clear

Table 2: Extracted quotes from participant answers to open response questions

B ADDITIONAL RESULTS

B.1 Task-Wise Evaluation Metrics

In Sec. 5.2 we examined the significance of effects for objective understanding disaggregated by the recourse-related task area (Sec. 4.3). Presented below are the understanding scores for each task area (range 0-4 for each), as well as the corresponding disaggregated response times for each question type in minutes.

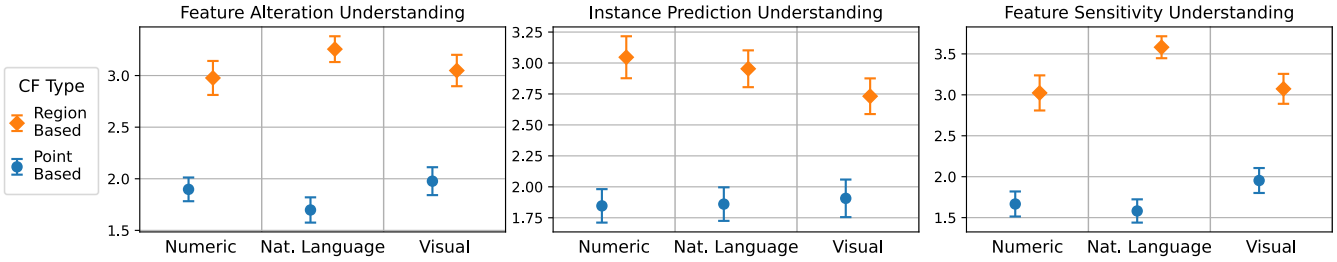


Figure 5: Mean and standard error of task-wise understanding for the six configurations of counterfactual explanation

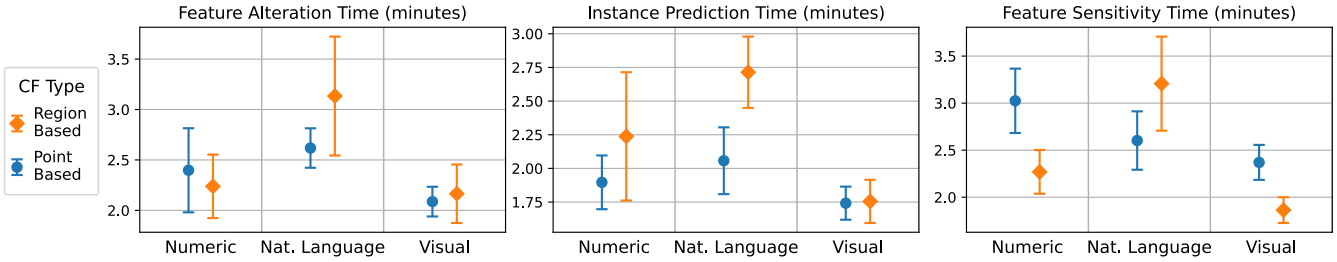


Figure 6: Mean and standard error of task-wise response time for the six configurations of counterfactual explanation

B.2 Participant Demographics

Tab. 3 shows a summary of the demographics of the analyzed participants using questions from [21]. Note that as all questions were optional and we allowed multiple responses per participant for race, the sum of each variable may not exactly match the total sample size. We collected age in brackets to preserve anonymity and gender information as an open response which was parsed into *nonbinary*, *man*, or *woman*. We group the four responses for *some high school* with the 23 for *high school diploma or equivalent* to create the *high school or less* category and merge the four from *applied or professional degree* into *other*.

Age		Education		Race		Gender	
18-24	26	High school or less	27	American Indian or Alaska Native	4	Nonbinary	3
25-34	86	Some college, no degree	54	Asian	16	Man	115
35-44	71	Associate degree	26	Black or African American	33	Woman	124
45-54	39	Bachelor's degree	105	Hispanic, Latino or Spanish Origin	20		
55-64	19	Master's degree	33	Middle Eastern or North African	1		
65-74	7	Doctorate degree	1	Native Hawaiian or Other Pacific Islander	1		
≥ 75	2	Other	5	White	202		

Table 3: Self reported demographic data of the 252 participants

B.3 Individual Factors

As literacy with computer systems or financial data may affect a participant's understanding, we adapted three agreement statements from existing research [6, 57] to measure each concept on a 6-point scale. We further directly ask participants to report their familiarity in each concept from *no knowledge* to *a lot of knowledge* on 4-point scales as in [6]. We normalized and summed the responses to create separate *technical literacy* and *financial literacy* scores, each ranging from 0-25. As perceptions of the appropriateness of the use of automated decision systems may impact user behavior, we also collected an *AI Sentiment* score by asking participants to indicate their agreement with *I believe it's okay for algorithms to be used to make important decisions* on a 6-point scale. See Appx. C.4 for a full list of these questions.

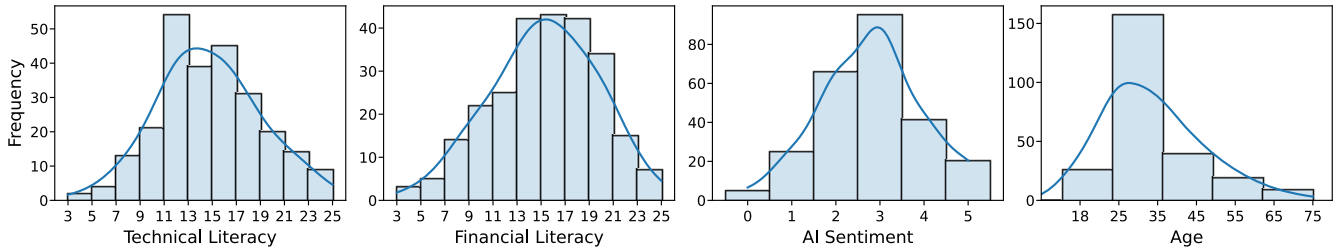


Figure 7: Distribution of individual factors and age for the analyzed participants (KDE smoothed, Scott's rule factor=1.3)

A multiple linear regression using *technical literacy*, *financial literacy*, and *AI sentiment* (Fig. 7) to predict objective understanding or *response confidence* reveal no significant effects. However, the same factors predicting *subjective understanding* reveals a potential effect

of *technical literacy* ($p = 0.0492, \beta = 0.1821$), indicating that participants who are more familiar with general technology may perceive themselves as more capable of understanding the automated decision system.

C SURVEY MATERIALS

C.1 Objective Understanding Questions

Below is the full text of the objective understanding questions as adapted from [4, 7]. Feature alteration wording and choices come directly from [7] with values for each option chosen from the explanation – i.e., the explanation altered both features and one value was chosen to match the counterfactual and the other not. For one of the four alteration questions we chose both values to not match, making "neither" the correct option. For instance prediction questions, the options were created by changing some features of the instance to not-match the counterfactual values. This ensured the new instances remained relevant to the explained instance. Values were chosen such that two of four prediction questions were Approved and the other two Rejected. New instances for feature sensitivity were generated by altering the instance to match the explanation then "moving" one feature-value to be near the end of the Approved range.

Question Type 1: Feature Alteration. Given the following explanation information [Explanation UI]. Which change is most likely to get this application approved?

(1) [Decreasing <feature i> from <value 1> to <value 2>
(2) Increasing <feature j> from <value 3> to <value 4>
(3) Neither would increase the chance of approval

Question Type 2: Instance Prediction. Given the following explanation information [Explanation UI]. Consider an applicant with the following profile

Attribute	Value
Applicant Income	$<value>
Coapplicant Income	$<value>
Loan Amount	$<value>
Loan Term	<value> Days

How would the algorithm categorize this applicant?

(A) Very likely to be rejected
(B) Somewhat likely to be rejected
(C) Somewhat likely to be accepted
(D) Very likely to be accepted

Question Type 3: Feature Sensitivity. Imagine that you applied for a loan and were REJECTED with the following explanation [Explanation UI]. You have now changed your application to the following values and been APPROVED for a loan

Attribute	Value
Applicant Income	$<value>
Coapplicant Income	$<value>
Loan Amount	$<value>
Loan Term	<value> Days

Which attribute of your new application, if changed by a small amount is most likely to result in a rejection?

(A) Applicant Income
(B) Coapplicant Income
(C) Loan Amount
(D) Loan Term

C.2 Subjective Understanding Questions

Metric	Please indicate how much you agree with the following statements.
Subj. Understanding	Explanations of the algorithm are easy to understand
Subj. Understanding	Given an explanation, I can reliably predict how the algorithm will behave
Subj. Understanding	Explanations of the algorithm help me understand how the approval decision is made
Subj. Understanding	Explanations of the algorithm help me increase the likelihood of getting my application approved
Subj. Understanding	I understand the criteria for loan approval
Responses	*strongly disagree, disagree, somewhat disagree, agree, strongly agree.*

C.3 Additional Metric Questions

Metric	Please indicate how much you agree with the following statements.
Trust	I trust the decisions made by the algorithm
Responses	*strongly disagree, disagree, somewhat disagree, agree, strongly agree.*

Satisfaction	Overall, how satisfied are you with the explanations provided for obtaining loan approval?
Responses	*not satisfied, a little satisfied, somewhat satisfied, satisfied, and highly satisfied.*

C.4 Individual Factors Questions

Individual Factor	Please indicate how much you agree with the following statements.
Technical Literacy	I am confident using computers
Technical Literacy	I understand how Amazon recommends products for me to choose
Technical Literacy	I can make use of computer programming to solve a problem.
Financial Literacy	I understand how my credit score is calculated
Financial Literacy	I understand how to file my own taxes
Financial Literacy	I feel capable of making important financial decisions
AI Sentiment	I believe it's okay for algorithms to be used to make important decisions
Responses	*strongly disagree, disagree, somewhat disagree, agree, strongly agree.*

Individual Factor	Question
Technical Literacy	How much programming knowledge do you have?
Technical Literacy	How much knowledge of computer algorithms do you have?
Responses	*no knowledge, a little knowledge, some knowledge, a lot of knowledge.*

Financial Literacy	How familiar are you with financial data?
Financial Literacy	How familiar are you with the credit approval processes such as making decisions about approving credit cards, loans, and mortgages?
Responses	*not familiar, a little familiar, very familiar, extremely familiar.*

C.5 Scenario Introduction

The following information was used to introduce all participants to the loan applicant scenario and explanation UI.

Introduction. Here we introduce information you will need to answer questions in this survey. Please read carefully as you can later earn bonus payment for correct answers.

Scenario. ACME Bank has developed a computer algorithm to automatically process loan applications. The algorithm automatically decides if a loan application should be APPROVED or REJECTED. Which of the following statements is TRUE?

- The algorithm is a set of rules that bank staff follow to manually make application decisions
- The algorithm is a computer program that automatically makes application decisions
- The algorithm is a computer program that randomly generates a number

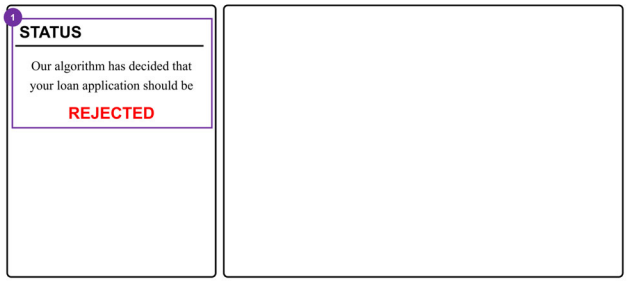
Applicant Information. The algorithm learns from historical data to decide if an application should be APPROVED. For example, the algorithm may approve an applicant if their profile is similar to those of previously APPROVED applicants. The algorithm uses the following attributes to make approval decisions.

Attribute	Details
Applicant Income	The primary applicant's total monthly income in dollars
Coapplicant Income	The total monthly income of the loan applicant's cosigners (such as a friend, partner, or parent) in dollars
Loan Amount	The total loan amount in dollars
Loan Term	The duration in days that the loan will be repaid over

Which of the following statements about the algorithm is FALSE?

- The algorithm learns from historical loan data
- The algorithm uses an applicant's income as part of its decision making
- The algorithm randomly decides which applicant to approve

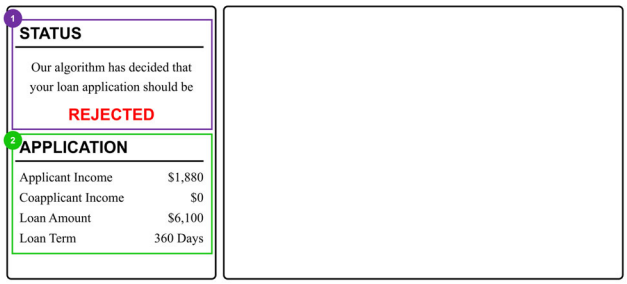
Explanation Tool. Imagine that you are a loan applicant who has applied for a loan. Your goal is to understand how the algorithm works with the explanation tool below. (Area 1) shows whether the algorithm has APPROVED or REJECTED your loan.



What decision did the algorithm make for your application?

- APPROVE
- REJECT

(Area 2) shows the values for each attribute of your application.



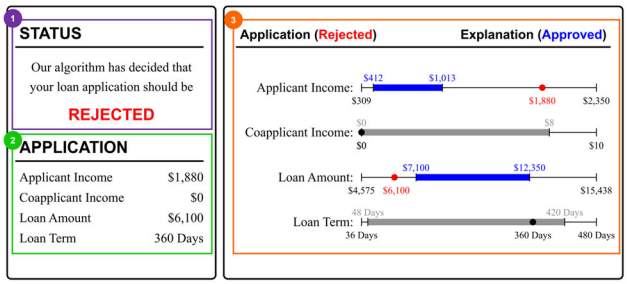
Which of the following matches the Applicant Income?

- $0
- $1,880
- $6,100
- Not shown

C.6 Explanation Training

The following information was presented to participants right after the Introduction. The explanation images shown were customized to match the each explanation configuration for each group.

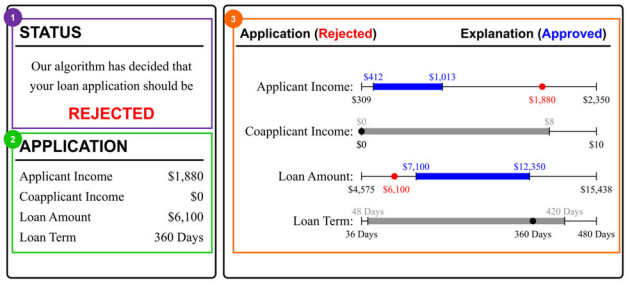
(Area 3) includes an explanation of the algorithm's decision.



The explanation tool shows changes you could make to get your application APPROVED

- True
- False

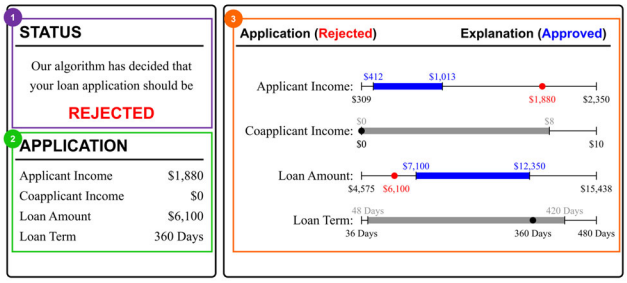
In (Area 3) values for attributes which must change are shown in Red. The proposed new values are shown in Blue. Values for unchanged attributes are shown in Grey.



The explanation above suggests changing the Loan Amount

- True
- False

The explanation below indicates your application was REJECTED, but would be APPROVED if you decrease your Applicant Income AND increase your Loan Amount. This explanation leaves Coapplicant Income and Loan Term unchanged. For this survey please do not consider your personal preferences or prior expectations about changing any given attribute



You should consider your personal preferences or expectations when considering changed attributes

- True
- False