Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on Human-Centred AI

Reassessing Evaluation Functions in Algorithmic Recourse: An Empirical Study from a Human-Centered Perspective

Tomu Tominaga[1] , **Naomi Yamashita**[2] , **Takeshi Kurashima**[1]
[1]NTT Human Informatics Laboratories, NTT Corporation
[2]NTT Communication Science Laboratories, NTT Corporation
{tomu.tominaga, takeshi.kurashima}@ntt.com, naomiy@acm.org

Abstract

In this study, we critically examine the foundational premise of *algorithmic recourse* – a process of generating counterfactual action plans (i.e., recourses) assisting individuals to reverse adverse decisions made by AI systems. The assumption underlying algorithmic recourse is that individuals accept and act on recourses that minimize the gap between their current and desired states. This assumption, however, remains empirically unverified. To address this issue, we conducted a user study with 362 participants and assessed whether minimizing the distance function, a metric of the gap between the current and desired states, indeed prompts them to accept and act upon suggested recourses. Our findings reveal a nuanced landscape: participants' acceptance of recourses did not correlate with the recourse distance. Moreover, participants' willingness to act upon recourses peaked at the minimal recourse distance but was otherwise constant. These findings cast doubt on the prevailing assumption of algorithmic recourse research and signal the need to rethink the evaluation functions to pave the way for human-centered recourse generation.

1 Introduction

As artificial intelligence (AI) has played an active role in high-stake decision making, algorithmic recourse has gained significant attention as a key explainable AI (XAI) technology. To help individuals who receive unfavorable decisions from AI systems, algorithmic recourse provides a counterfactual action plan, called *recourse*, for them to flip the decision and reach a favorable outcome [Karimi *et al.*, 2022] such as "your loan application would be approved if you increased the annual income by $10K and changed the educational background from bachelor to master". Its ultimate goal is to offer recourses for unfavorable decisions so that individuals can accept and act upon them. .

To generate such optimal recourses, algorithmic recourse research has addressed the challenge of identifying the counterfactual sample that has minimal disparity with the target individual. Here, counterfactual samples refer to user samples chosen to contrast with target individuals within recourses, selected from among those who receive favorable decisions from the AI system. The disparity between target and counterfactual individuals is often evaluated with distance functions that quantify the feature differences between them using norm-based measures [Verma *et al.*, 2020]. This task is typically formulated as an optimization problem [Wachter *et al.*, 2018], with numerous recent proposals for technical solutions [Karimi *et al.*, 2022; Verma *et al.*, 2020].

Such technological advancements rest upon a crucial assumption that individuals accept and act on recourses generated through the minimization of the evaluation functions. To date, the evaluation functions have been developed to capture the simplicity of explanations and the effortlessness of action plans in the suggested recourses. Concerning simplicity, researchers presume that individuals prefer recourses with fewer changes suggested. This is inspired by Miller's insights [Miller, 2019], emphasizing people's preference for simpler explanations with fewer causes cited [Thagard, 1989; Read and Marcus-Newhall, 1993]. For effortlessness, acknowledging that transitioning from the current to the desired state incurs costs, they posit that recourses with minimal suggested changes are easier for individuals to implement [Ustun *et al.*, 2019; Karimi *et al.*, 2022]. Hence, it has been presumed that recourses minimizing proposed changes assessed by evaluation functions are optimal for individuals.

However, little empirical evidence supports the underlying assumption that individuals indeed accept and act upon the recourses minimizing the evaluation functions. As evidenced in persuasive technology research [Fogg, 1998], reshaping human attitudes and behaviors through computational approaches is challenging, suggesting that the recourse generation process is more nuanced than initially assumed. This lack of validation for the aforementioned assumption can be attributed to the XAI community's focus on technological advancements, often overlooking the importance of rigorous evaluation through user studies [Keane *et al.*, 2021; Adadi and Berrada, 2018]. To achieve robust scientific progress towards human-centered algorithmic recourse generation, it is imperative to confirm the validity of the evaluation function – the core principle of recourse generation that acts as the objective function in the optimization problem. This validation

Supplementary materials for this paper are available in the extended version at: https://doi.org/10.48550/arXiv.2405.14264

should be achieved through user studies [Barocas *et al.*, 2020] and psychologically grounded [Keane *et al.*, 2021].

In this paper, we aim to provide empirical evidence for the premise underlying algorithmic recourse by addressing the following research question.

Research Question: *How does the recourse distance, as quantified by evaluation functions, affect individuals' willingness to accept and act upon recourses?*

This study employs L_0 and L_1 norms, referred to as *sparsity* and *proximity* respectively, as the evaluation functions in this study due to their fundamental roles in algorithmic recourse research [Karimi *et al.*, 2022; Verma *et al.*, 2020].

To investigate the research question, we conducted an online study with 362 participants. Our experiment involved a real-world car loan application scenario, as financial assessments such as loan screening or credit evaluation are one of the paramount themes in algorithmic recourse research [Kirfel and Liefgreen, 2021; Wang *et al.*, 2023]. To enhance the experiment's authenticity, we recruited participants genuinely interested in a car loan. We systematically devised various recourses tailored to participants' profiles, controlling the recourse distance using the evaluation functions, and evaluated their inclinations to accept and act upon these recourses.

Contributions. This study presents a unique contribution by validating and challenging the foundational assumption of algorithmic recourse research through user experiments. Notably, it provides empirical evidence that participants' willingness to accept recourses was not affected by the recourse distance. Furthermore, the likelihood of acting on recourses was maximized at minimal recourse distance but remained constant elsewhere. These findings underscore the urgent need to revisit the evaluation functions used in algorithmic recourse. Additionally, this study enriches the field by outlining future research directions crucial for developing a human-centered approach to algorithmic recourse generation AI.

2 Related Work

2.1 Algorithmic Recourse Generation

In the rapidly evolving research field, the definition of algorithmic recourse varies [Joshi *et al.*, 2019; Ustun *et al.*, 2019; Venkatasubramanian and Alfano, 2020], but its overarching objective is consistent: to help individuals subjected to unfavorable AI decisions understand the rationale behind them [Joshi *et al.*, 2019; Wachter *et al.*, 2018] and subsequently achieve favorable outcomes [Karimi *et al.*, 2021; Karimi *et al.*, 2020a]. Drawing on the concept that individuals favor simpler explanations [Miller, 2019] and are more inclined to undertake recourse that proposes smaller changes [Karimi *et al.*, 2022], researchers have tackled the following optimization problem to identify counterfactual samples for constructing recourses: given a fixed predictive model $h : X \to \{-1, +1\}$ (e.g., -1 is "rejection" and $+1$ is "approval") with an N-dimension feature space $X = X_1 \times \cdots \times X_N$ and its subspace $S(X) \subseteq X$ determined by arbitrary conditions on the features, find a counterfactual sample $x' \in X$ of an input sample $x \in X$ ($h(\boldsymbol{x}) = -1$).

$$\boldsymbol{x'} \in \underset{\boldsymbol{x'} \in S(X)}{\operatorname{argmin}}\, d(\boldsymbol{x}, \boldsymbol{x'}) \text{ subject to } h(\boldsymbol{x'}) \neq h(\boldsymbol{x}) \quad (1)$$

In this task, $d(\cdot, \cdot) : X \times X \to \mathbb{R}_{\geq 0}$ is a distance function to measure the distance between $\boldsymbol{x}$ and $\boldsymbol{x'}$. Among various types of distance functions, the most fundamental ones are norm-based measures (e.g., L_1 norm) [Verma *et al.*, 2020].

To date, many studies have predominantly focused on enhancing technical performance such as capturing feature distributions [Kanamori *et al.*, 2020; Poyiadzi *et al.*, 2020], incorporating causal knowledge among features into constraints [Karimi *et al.*, 2021; Karimi *et al.*, 2020b], making coherent and diverse explanations [Russell, 2019; Mothilal *et al.*, 2020], or computing the order of actions [Kanamori *et al.*, 2021; Naumann and Ntoutsi, 2021]. While the optimization problem inherently suggests that minimizing the distance functions will lead algorithmic recourse to produce recourses that are both acceptable and actionable for individuals, empirical evidence supporting this notion is scarce. This study aims to address this issue by conducting user experiments to verify the aforementioned assumption.

2.2 Human Evaluation of Algorithmic Recourse

Since the lack of human-centered evaluation for counterfactual suggestions was identified [Keane *et al.*, 2021; Verma *et al.*, 2020], there has been a gradual increase in user studies [Förster *et al.*, 2020; Förster *et al.*, 2021; Kanamori *et al.*, 2022; Kirfel and Liefgreen, 2021; Rawal and Lakkaraju, 2020; Singh *et al.*, 2023; Warren *et al.*, 2023; Yacoby *et al.*, 2022]; however, studies investigating individuals' willingness to accept and act on counterfactual suggestions remain limited. Among the limited research, an exploratory study using an interactive recourse generation system has shown that recourses with fewer and more controllable changes were actionable [Wang *et al.*, 2023]. Additionally, players in a simulation game were found to adopt strategies that mirrored their previously employed tactics [Kuhl *et al.*, 2022].

However, these studies do not use evaluators' profile data to generate recourses [Wang *et al.*, 2023; Kirfel and Liefgreen, 2021] and ignore real-world contexts [Kuhl *et al.*, 2022], thereby failing to accurately replicate realistic scenarios where people encounter adverse AI decisions. Moreover, these studies focus more on generating recourses rather than evaluating users' preferences. Our study aims to bridge this gap by constructing recourses based on participants' profile data within real-world contexts, offering a more authentic evaluation. This approach not only confirms the validity of evaluation functions but also enhances our understanding of users' propensity to accept and act on recourses.

3 Experiment

We carried out an experiment to assess how individuals' attitudes towards accepting and acting on recourses change with the recourse distance. To simulate real-world scenarios where individuals encounter unfavorable AI decisions, we developed a scenario involving car loan applications and selectively recruited participants who fit this specific context. In this scenario, participants submitted their profile information for the application, received a rejection notification based on the assessment, and then rated their willingness to accept and act upon the suggested recourses.

3.1 Experimental Setup

Hypothetical Scenario for Recourse Provision

We crafted a hypothetical scenario where participants were asked to imagine themselves applying for a car loan at a financial institution. There are two reasons for choosing the scene of car loans. First, finance is one of the most promising domains in algorithmic recourse research [Ustun *et al.*, 2019; Karimi *et al.*, 2020a; Wang *et al.*, 2023], with loan or credit screening being a predominant case [Kirfel and Liefgreen, 2021; Wang *et al.*, 2023]. Second, among items financed through loans such as education loans, car loans, and mortgages, cars are typically well-known, relatively expensive, and are actively owned and used by the buyer; therefore, the car loan application scenario is likely to prompt participants to easily imagine themselves in the situation, think it as a high-stakes decision, and relate it to their real-life context when evaluating recourses.

Scenario Design

The participants engaged in the following car loan application scenario:

> *The participant wants to take out a two-year car loan equivalent to one-third of their annual income. They visit a financial institution to obtain the loan and are asked to submit their profile data for screening. The screening process involves an AI system that determines applicants' eligibility based on an extensive database of customer information held by the financial institution. After submitting their profile data, they are informed of the rejection of their loan application shortly thereafter. To find out why the application was rejected and what action is required to secure approval, they will view the recourses generated by the AI system, derived from their profile data and the financial institution's customer database.*

Here, the loan amount in this experiment is relative to the participants' annual income rather than a fixed amount, with a repayment term fixed at two years. This scenario design ensures uniformity in the repayment burden across participants, thereby minimizing its influence on their evaluation of the likelihood of acting upon the suggested recourses.

We constructed recourses based on the assumption that the financial institution has a fixed rule of rejecting any loan applications exceeding one-quarter of the applicant's annual income. Therefore, it should be noted that all the participants experienced rejection in this experiment. This rule was not disclosed to the participants throughout the experiment.

3.2 Recourse Construction

To enable participants to evaluate recourses across diverse distances, we selected five counterfactual samples for each input sample (i.e., a participant's profile data). To accomplish this, we pre-collected 4057 profile data points to serve the counterfactual samples, representing "an extensive database of customer information held by the financial institution" in the scenario (see Section A in supplementary materials for further details). Using this counterfactual sample pool, we constructed recourses as follows.

Conditions

A counterfactual sample $\boldsymbol{x'}$, used for generating a recourse $\boldsymbol{\delta}$ given an input sample $\boldsymbol{x}$, must meet the following conditions: (1) the annual income of $\boldsymbol{x'}$ is 4/3 times greater than that of $\boldsymbol{x}$ and (2) $\boldsymbol{x'}$ and $\boldsymbol{x}$ follow the constraints in Table 1. The first one echoes the fixed decision rule of the financial institution in the experimental scenario. Under the fixed rule, the car loan applications are accepted if the annual income is 4 times or more of the loan amount. In the scenario, the loan amount applied is one third of the annual income. We then set the first condition to select counterfactual samples for recourse customization.

The second condition relates the feature constraints. As described in Table 1, these constraints aim to exclude impossible or extremely rare feature changes. Examples include raising the educational degree (#3), work position (#5), service years (#6), and management career (#7), but it is impossible (extremely rare) to lower them. It is also impossible to remove one's experience and skills such as the job change experience (#11), overseas working experience (#12), overseas study experience (#13), and best test score (#14) from oneself. Other than the above features, we do not impose any constraints. These constraints enable us to generate recourses changing features within a theoretically modifiable range.

Distance Metrics

For distance metrics as the evaluation functions of recourse $\boldsymbol{\delta}$, we adopted the L_0 norm as the sparsity and the L_1 norm as the proximity as follows. Given input sample $\boldsymbol{x}$ and its counterfactual sample $\boldsymbol{x'}$, we firstly computed

$$\delta_i = \begin{cases} \mathbb{I}[x'_i \neq x_i] & (\text{if } i\text{-th feature is categorical}) \\ |x'_i - x_i|/M_i & (\text{otherwise}) \end{cases} \tag{2}$$

Here, $\mathbb{I}[\cdot]$ is the indicator function. M_i is the range of the i-th feature change to normalize the distances of all features to scales ranging from 0 to 1. We then calculated

$$\text{sparsity} = \sum_i \mathbb{I}[\delta_i \neq 0], \ \text{proximity} = \sum_i \delta_i. \tag{3}$$

We adopted these metrics due to their foundational significance and widespread acceptance in the research community [Verma *et al.*, 2020], despite the presence of several other alternative metrics [Ustun *et al.*, 2019; Wachter *et al.*, 2018]. Note that these two metrics exhibit lower values as the recourse is more sparse or proximate, following the technical definitions in prior research [Karimi *et al.*, 2022; Verma *et al.*, 2020; Wachter *et al.*, 2018].

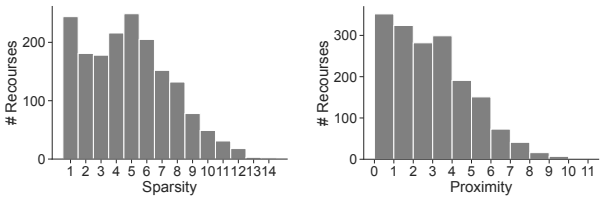
Selection Process

From the subset of the counterfactual sample pool extracted by the above conditions, we selected five counterfactual samples for each participant for recourse customization and evaluation using the following strategy: (1) select one of the sparsest samples, (2) select one of the most proximate samples, and (3) randomly select three additional samples.

Given that the main objective of our experiment is to evaluate participants' responses to recourses across a spectrum of distances, we employed the first and second strategies to select short-distance recourses, while the third strategy was used to introduce a broader variety of distances (Figure 1).

#	Item (Feature)	Option	Constraints
1	Residential prefecture	1. Tokyo / 2. Other than Tokyo	$x'_1 \geqq x_1$
2	Type of residence	1. Owned house / 2. Rental housing	$x'_2 \geqq x_2$
3	Educational background	1. High school / 2. Junior college / 3. University (bachelor) / 4. Graduate school (master) / 5. Graduate school (doctor)	$x'_3 \geq x_3$
4	Workplace	1. Private company / 2. Public institution	$x'_4 \geqq x_4$
5	Position	1. Employee / 2. Supervisor / 3. Section Head / 4. Section Chief / 5. Assistant General Manager / 6. Manager / 7. General Manager / 8. Executive Director / 9. Senior Executive Director / 10. President	$x'_5 \geq x_5$
6	Service years	1. 0-1 year / 2. 1-3 years / 3. 3-5 years / 4. 5-10 years / 5. 10-20 years / 6. 20- years	$x'_6 \geq x_6$
7	Management career	1. No / 2. 0-1 year / 3. 1-3 years / 4. 3-5 years / 5. 5-10 years / 6. 10-20 years / 7. 20- years	$x'_7 \geq x_7$
8	Working hours per day	1. 0-2 hours / 2. 2-4 hours / 3. 4-6 hours / 4. 6-8 hours / 5. 8-10 hours / 6. 10-12 hours / 7. 12- hours	$x'_8 \leqq x_8$
9	Teleworking hours per day	1. 0-2 hours / 2. 2-4 hours / 3. 4-6 hours / 4. 6-8 hours / 5. 8-10 hours / 6. 10-12 hours / 7. 12- hours	$x'_9 \leqq x_9$
10	The number of side jobs	1. No / 2. 1 job / 3. 2 jobs / 4. 3 jobs / 5. 4 jobs / 6. 5- jobs	$x'_{10} \leqq x_{10}$
11	Job change experience	1. No / 2. Yes	$x'_{11} \geq x_{11}$
12	Overseas working experience	1. No / 2. Yes	$x'_{12} \geq x_{12}$
13	Overseas study experience	1. No / 2. Yes	$x'_{13} \geq x_{13}$
14	TOEIC* Best score	1. No / 2. 10-400 / 3. 400-495 / 4. 500-595 / 5. 600-695 / 6. 700-795 / 7. 800-895 / 8. 900-990	$x'_{14} \geq x_{14}$
15	Facebook app use	1. No / 2. Yes	$x'_{15} \leqq x_{15}$
16	LinkedIn app use	1. No / 2. Yes	$x'_{16} \leqq x_{16}$

Table 1: Profile data included in recourses. x'_i and x_i are i-th elements of an input and counterfactual samples (e.g., $x'_5 = 2$ means that the counterfactual sample's Position is a Supervisor). The constrains column describes feature conditions of a counterfactual sample $\boldsymbol{x'}$ to be selected for a suggested recourse given an input sample $\boldsymbol{x}$.
*The Test of English for International Communication (TOEIC®) Listening & Reading Test (https://www.iibc-global.org/english/toeic/test/lr.html), one of the most popular English language proficiency tests in Japan. The score is in 5-point increments from 10 to 990.



(a) Sparsity (b) Proximity

Figure 1: Distributions of distance metrics of selected recourses.

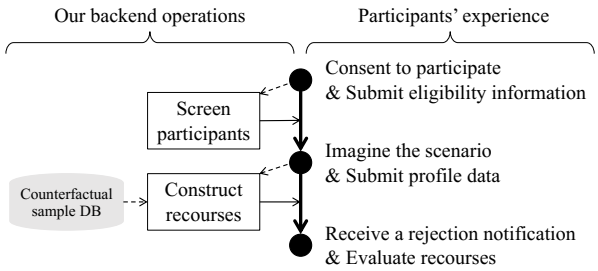


Figure 2: Overview of the experiment process. Black circles indicate steps taken by participants, white boxes represent interventions from us to participants, and thin dashed arrows depict data flow.

3.3 Experimental Procedure

Participants first took a screening survey to assess their study eligibility, followed by filling out a consent form. They then engaged with a hypothetical car loan scenario and submitted their profile data. Finally, they received a rejection notification along with five recourses and evaluated each through questionnaires (Figure 2). Details on the questionnaires can be found in Section B of the supplementary materials.

The experiment was conducted from July 7th to August 10th, 2023. Participants received a compensation of 100 JPY for their participation.

Participant Recruitment and Screening

A total of 362 Japanese participants were recruited through a recruiting company in Japan, meeting the following criteria: (1) employment in private companies or public institutions (full-time or part-time), (2) interest in applying for a car loan, (3) no existing loans, and (4) an annual income less than JPY 10M. The first criterion targeted paid workers, excluding students, to ensure participants had stable incomes. The second

and third criteria were designed to select participants interested in and financially capable of considering a car loan. For instance, those with existing loans might lack the financial flexibility to pursue new ones, leading to reluctance in engaging with new financial commitments. The fourth criterion facilitated the selection of five counterfactual samples for each participant from the counterfactual sample pool (see Section A in supplementary materials for further details).

Profile Data Submission

Following the screening process, participants provided their profile data relevant to the experimental scenario. The profile data items are shown in Table 1. We selected items covering basic demographics (#1-3), current employment status (#4-10), professional experience and skills (#11-14), and human connections (#15, #16). These were chosen to mirror the attributes commonly included in machine learning datasets for financial credit judgments [Hofmann, 1994; Yeh, 2016; Becker and Kohavi, 1996] and information used in a typical resume [Brown and Campion, 1994; Cole *et al.*, 2007], assuming these reflect critical screening items for evaluating applicants' creditworthiness and employment stability.

Here, we deliberately avoided profile data categorized as immutable features - those that individuals cannot change (e.g., gender, birthplace, or ethnicity) [Kirfel and Liefgreen, 2021] because recourses including such features sometimes acceptable but always unactionable [Karimi *et al.*, 2022].

We also excluded annual income from the profile data to maintain the focus of our scenario-based experiment. In this experiment, participants' applications were declined due to insufficient annual income relative to the requested loan amount. As such, if annual income is included as a profile data item, recourses inevitably advise its adjustment. This could narrow the focus of our evaluation to participants' reactions to changes in annual income, rather than their perceptions of recourse sparsity. To address this and ensure that a diverse range of features are considered in 1-sparsity recourses, annual income is omitted from the profile data.

Recourse Evaluation

After the profile data submission, participants were instructed to wait two weeks for the review. During this period, we developed customized recourses (Section 3.2), refined the questionnaire for evaluating these recourses, and sent it as a review report for each participant. Upon receiving this report, participants were informed that their applications had been rejected. They were then asked to evaluate the recourses generated by the AI system as outlined in Table 2.

To assess the participants' evaluations of recourses, we posed the following two questions: *"Is the AI system's plan a reasonable explanation for the rejection of your loan application?"* to gauge their propensity to accept recourses, and *"Would you carry out the plan presented by the AI system to obtain loan approval?"* to measure their willingness to act upon recourses. For both questions, responses were captured on a 7-point scale from *"Strongly no"* to *"Strongly yes"*. Participants were also asked to explain the reasons for their recourse evaluations through an open-ended question.

We also examined the immutability of changing features as suggested in the recourses. As noted in Section 3.3, immutable feature changes should be excluded from the recourses because they inherently make recourses unactionable. Given that the immutability of altering features depends on personal situations, we asked participants about this aspect to filter out impractical recourses from further analyses.

	☹ Current profile		☺ Ideal profile
...	...		...
Workplace	Private company	→	Private company
Position	Employee		Supervisor
...	...		...

Table 2: A partially excerpted example of a recourse presented to participants (see Section B3 in supplementary materials for a full example). The current profile is the participant's profile and the ideal profile is her/his counterfactual sample. All the items are displayed irrespective of modifications. Items recommended for a change are highlighted in green and indicated with an arrow.

4 Analysis

We collected a total of 1810 sets of recourse evaluation data. Among them, 72 had at least one immutable feature change. After removing these, we analyzed the remaining 1738 evaluations quantitatively and qualitatively.

4.1 Quantitative Analysis

To investigate the connections between individuals' willingness to accept or act on the recourses and recourse distance, we utilized generalized additive mixed models (GAMMs). This analytical approach combines the characteristics of generalized additive models and mixed-effect models to estimate the nonlinear dependencies between the objective and explanatory variables, accounting for individual subject-specific effects. We established GAMMs as follows, where Y represents the propensity to accept or act upon recourses, x denotes the recourse distance, and ϵ denotes the error term, using the parameters θ and the smoothing spline function $s(\cdot)$. Note that i denotes the observation, u represents the participant, and j shows the smoothing spline function.

$$\begin{aligned} Y_{iu} &= \theta_{0u} + \sum_{j=1}^{K} \theta_{ju} s_j(x_{iu}) + \epsilon_{iu} \\ \theta_{0u} &= \beta_{00} + b_{0u},\ b_{0u} \sim \mathcal{N}(0, \sigma_0^2) \\ \theta_{ju} &= \beta_{j0} + b_{ju},\ b_{ju} \sim \mathcal{N}(0, \sigma_j^2) \end{aligned} \tag{4}$$

Nesting the random effects b associated with individual participants within the intercept θ_0 and slope θ_j enables us to discern the fixed effects β of recourse distance. We fitted the aforementioned GAMMs using the mgcv package 1.9-0 421 in RStudio 2023.09.1+494.

4.2 Qualitative Analysis

The free-form survey responses added nuance to our quantitative results, shedding light to participants' perspectives on (non-)sparse/proximate recourses, as well as their rationales for favoring one over the other. Specifically, we examined

the responses to the open-ended survey question, *"Why did you make the evaluation you did?"*, following their ratings on propensity for acceptance and action, respectively.

5 Results

We formulated four distinct GAMMs based on combinations of the dependent variable (i.e., propensity to accept or to act upon recourses) and the explanatory variable (i.e., sparsity or proximity). Figure 3 illustrates the smoothing spline curves generated from applying these GAMMs to the experimental data. For a summary of statistics obtained from these models, please refer to Table S2 in the supplementary materials.

5.1 Propensity to Accept Recourses

Quantitative Results

We found that individuals' propensity to accept recourses is not associated with the distance metrics. As shown in the statistics from the GAMM detailed in Table S2, both sparsity and proximity did not exhibit a significant relationship with propensity for acceptance (Model-1: $F = 0.00$, $p = 1.00$; Model-2: $F = 0.89$, $p = 0.347$). Figure 3(a) illustrates the smoothing spline functions obtained from the GAMMs. Propensity to accept recourses remains unchanged across different values of distance metrics. This result challenges the foundational premise that sparser or more proximate recourses are more acceptable.

Qualitative Results

Analysis of the open-ended responses revealed that participants who deemed sparse or proximate recourses as unacceptable expressed skepticism about their failure to pass the car loan screening, especially when their profile data was almost identical to their counterfactual samples. For example, P006 commented *"Almost same"* when referring to a minor change (sparsity = 1; proximity = 0.167; Teleworking hours per day: 0-2 hours $\rightarrow$ 2-4 hours) and rated the recourse unacceptable (propensity for acceptance = 1). Similarly, P446 questioned *"That's it?"* for a slight adjustment (sparsity = 1; proximity = 0.40; Educational background: Junior college $\rightarrow$ University) and gave a low score for accepting the recourse (propensity for acceptance = 2). P706 reported *"Not particularly different compared to the current situation"* for a modification (sparsity = 2; proximity = 0.50; Position: Employee $\rightarrow$ Section Chief, Working hours per day: 10-12 hours $\rightarrow$ 8-10 hours) with a low acceptance rating (propensity for acceptance = 2).

Conversely, participants who rated non-sparse or non-proximate recourse acceptable often saw such challenging recourses as opportunities to introspect their capabilities and current status, attributing their loan application rejection to personal shortcomings. For example, P567 stated *"The career is too different"* for a significant shift (sparsity $= 10$; proximity $= 8.40$; e.g., Position: Employee $\rightarrow$ President), rating their acceptance high (propensity for acceptance $= 7$). P739 described *"I think it's a solid reason, so I don't blame it"* for a major transition (sparsity $= 10$; proximity $= 5.64$; e.g., Type of residence: Rental housing $\rightarrow$ Owned home, Service years: 3-5 years $\rightarrow$ 20- years), giving a high acceptance score (propensity for acceptance $= 6$). Lastly, P732 admitted *"Because I realize how much I'm not enough"* when faced with extensive changes (sparsity $= 10$; proximity $= 6.11$; e.g., Educational background: University $\rightarrow$ Graduate school (doctor), The number of side jobs: No $\rightarrow$ 2 jobs) and found it relatively acceptable (propensity for acceptance $= 5$).

5.2 Propensity to Act Upon Recourses

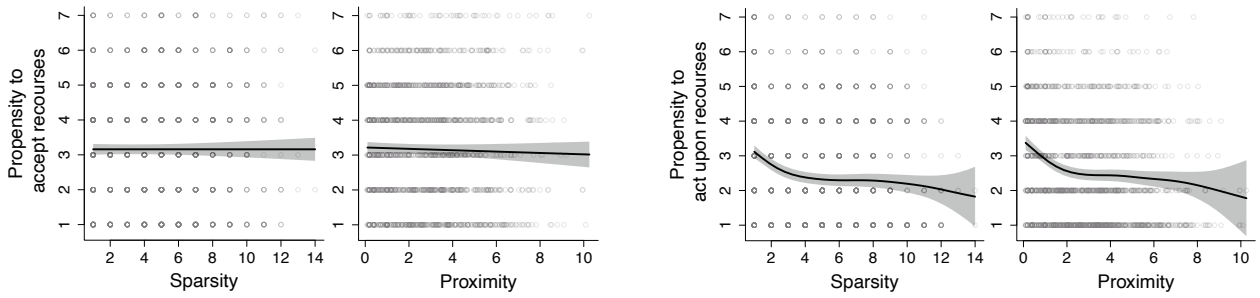
Quantitative Results

The spline terms for sparsity and proximity exhibited significant associations with propensity for action, as seen in Table S2 (Model-3: $F = 18.46$, $p < 0.001$; Model-4: $F = 20.81$, $p < 0.001$). In addition, the effective degrees of freedom (EDF) for the spline terms of sparsity and proximity were 4.25 and 4.32, respectively (Table S2). These results indicate a nonlinear rather than a strictly linear dependence between the propensity for action and the distance metrics. The smoothing spline functions of the distance metrics for the propensity for action are depicted in Figure 3(b), revealing that the propensity to take action exhibits a linear and monotonic decrease in the range of small distance metric values, but remains relatively constant across other ranges. In summary, the results indicate that participants were more willing to act on recourses that are both sparse and proximate, aligning with the assumption made in prior research. Notably, the propensity to act was highest at the minimal recourse distance but remained constant beyond a specific threshold.

Qualitative Results

Upon analyzing participants' explanations for their willingness to engage with sparse or proximate recourses, it became clear that they perceived these suggestions as readily executable tasks. For example, P674 stated *"Because I think I can tackle it right now"* about a simple adjustment (sparsity = 1; proximity = 0.14; TOEIC: No $\rightarrow$ 0-400; propensity for action = 5), and P718 mentioned *"If it was just working hours, I'd do it in a heartbeat"* regarding a minor change in working hours (sparsity = 1; proximity = 0.20; Working hours per day: 6-8 hours $\rightarrow$ 8-10 hours; propensity for action = 7). Additionally, the perception that these recourses were straightforward to implement boosted participants' motivation to act. For example, P294 expressed *"Because the contents are easy to change"* when commenting on a simple transition (sparsity = 1; proximity = 0.20; The number of side jobs: 1 job $\rightarrow$ No; propensity for action = 7), and P655 described *"All I have to do is wait a little"* regarding a minor increase in service years (sparsity = 1; proximity = 0.20; Service years: 0-1 years $\rightarrow$ 1-3 years; propensity for action = 6).

In contrast, when examining participant explanations on the assessment of non-sparse or non-proximate recourses that received low ratings of willingness to act, it was evident that participants cited increased burdens as a deterrent such as significant time and financial investment, a high volume of tasks to complete, or low perceived cost-effectiveness. For example, P401 expressed *"Because there's too much to do, it takes too much time and money"* when evaluating a major overhaul (sparsity = 11; proximity = 8.38; e.g., Residential prefecture: Other than Tokyo $\rightarrow$ Tokyo, Overseas working experience: No $\rightarrow$ Yes, LinkedIn: No $\rightarrow$ Yes; propensity for action = 2), P167 remarked, *"There are too many items to clear"*, in response to extensive requirements (sparsity =



(a) Smoothing spline curves of sparsity (left) and proximity (right) (b) Smoothing spline curves of sparsity (left) and proximity (right)

Figure 3: GAMM fits of the distance metrics to the propensity for acceptance (a) and action (b). The GAMMs include individual participant-specific effects as the nested random intercept and slopes. The error bars are 95% confidence intervals.

12; proximity =7.82; e.g., Type of residence: Rental housing → Owned house, Position: Supervisor → Section Head, TOEIC: 10-400 → 900-990; propensity for action = 1), and P468 stated *"I don't want to change various things just to apply for a loan"*, highlighting the reluctance to undertake numerous changes (sparsity = 11; proximity = 9.09; e.g., Educational background: High school → Graduate school (doctor), Management career: No → 5-10 years, The number of side jobs: No → 1 job; propensity for action = 1).

6 Discussion

Implications. Our quantitative and qualitative results challenge the underlying assumption of algorithmic recourse, showing that the evaluation functions fail to capture individuals' propensity to accept recourses. Contrary to prior studies' assumption, we observed a nuanced dynamic: individuals expressed skepticism towards AI decisions when faced with recourses that are extremely "close" to their current situation, while they tended to acknowledge their inadequacies and accept AI decisions when presented with recourses suggesting substantial changes. Furthermore, while the evaluation functions effectively capture the propensity for action up to a certain threshold, they fall short of representing it beyond that threshold. To generate optimal recourses by maximizing individuals' willingness to act on recourses, setting firm upper limits of evaluation functions is crucial. Although this approach has been hinted at in prior works [Van Looveren and Klaise, 2021; Pawelczyk *et al.*, 2020], our research validates its importance in enhancing user engagement with recourses.

Future directions. Designing evaluation functions that accurately reflect individuals' propensity to accept or act upon recourses remains a pressing challenge for advancing human-centered algorithmic recourse generation. The key to addressing this challenge lies in understanding the individual variance in recourse evaluation. Our analysis using GAMMs highlights that individual subject-specific effects explain the propensity to accept or act upon recourses (see statistics of the variable *uid* for all the models in Table S2). This finding points to the importance of individual differences in shaping evaluation functions, which is overlooked in prior studies. Future research should uncover key factors driving the individual differences and integrate them into evaluation functions. By doing so, it would be possible to create evaluation functions that adapt to the nuanced acceptance/action propensity of each individual, enabling the generation of recourses tailored to each person's preferences and circumstances.

Limitations. Our experiment chose a car loan application scenario because financial screening is a paramount theme in recourse research. However, further investigation is needed to ascertain if the observed results generalize across different contexts. Additionally, our participant pool was exclusively Japanese, highlighting the need for future studies to include more diverse demographics spanning various countries, languages, and cultures. This is especially critical in light of our discussions on individual differences, underscoring the importance of broadening the scope to ensure the applicability of our findings across diverse global populations.

7 Conclusion

Our empirical findings challenge the core assumption of algorithmic recourse generation research by demonstrating that the conventional evaluation functions fail to effectively reflect individuals' acceptance and action propensities. This insight underscores the pressing need for a paradigm shift in the design of evaluation functions for the next generation of human-centered algorithmic recourse AI. Specifically, it advocates for incorporating adaptive adjustments of evaluation functions to accommodate individual differences. Such a refined approach holds promise for delivering recourses more personalized and beneficial to individual users, significantly improving the efficacy of recourse generation. We hope this research serves as a foundation for future scientific advancements and technological developments in XAI research.

Ethical Statement

This study involving human subjects was reviewed and approved by the external ethics review committee, the Institutional Review Board of Public Health Research Foundation[1] (the protocol number: PHRF-IRB 23F0002). All the procedures were performed in accordance with the guidelines.

[1]https://www.phrf.jp/rinri/ (only in Japanese)

References

[Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[Barocas *et al.*, 2020] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

[Becker and Kohavi, 1996] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[Brown and Campion, 1994] Barbara K. Brown and Michael A. Campion. Biodata phenomenology: Recruiters' perceptions and use of biographical information in resume screening. *Journal of Applied Psychology*, 79(6):897–908, 1994.

[Cole *et al.*, 2007] Michael S. Cole, Robert S. Rubin, Hubert S. Feild, and William F. Giles. Recruiters' Perceptions and Use of Applicant Résumé Information: Screening the Recent Graduate. *Applied Psychology*, 56(2):319–343, 2007.

[Fogg, 1998] BJ Fogg. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 225–232, 1998.

[Förster *et al.*, 2020] Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. Evaluating Explainable Artifical Intelligence – What Users Really Appreciate. In *European Conference on Information Systems (ECIS)*, pages 1–18, 2020.

[Förster *et al.*, 2021] Maximilian Förster, Philipp Hühn, Mathias Klier, and Kilian Kluge. Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 1274–1283, 2021.

[Hofmann, 1994] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[Joshi *et al.*, 2019] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *arXiv preprint arXiv:1907.09615*, 2019.

[Kanamori *et al.*, 2020] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization Kentaro. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2855–28, 2020.

[Kanamori *et al.*, 2021] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike, Kento Uemura, and Hiroki Arimura. Ordered Counterfactual Explanation by Mixed-Integer Linear Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11564–11574, 2021.

[Kanamori *et al.*, 2022] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 1846–1870, 2022.

[Karimi *et al.*, 2020a] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-Agnostic Counterfactual Explanations for Consequential Decisions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 895–905, 2020.

[Karimi *et al.*, 2020b] Amir Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. In *Advances in Neural Information Processing Systems*, 2020.

[Karimi *et al.*, 2021] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[Karimi *et al.*, 2022] Amir Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5), 2022.

[Keane *et al.*, 2021] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4466–4474, 2021.

[Kirfel and Liefgreen, 2021] Lara Kirfel and Alice Liefgreen. What If (and How ...)? - Actionability Shapes People's Perceptions of Counterfactual Explanations in Automated Decision-Making. In *ICML-21 Workshop on Algorithmic Recourse*, 2021.

[Kuhl *et al.*, 2022] Ulrike Kuhl, André Artelt, and Barbara Hammer. Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2125–2137, 2022.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[Mothilal *et al.*, 2020] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[Naumann and Ntoutsi, 2021] Philip Naumann and Eirini Ntoutsi. Consequence-Aware Sequential Counterfactual Generation. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 682–698, 2021.

[Pawelczyk *et al.*, 2020] Martin Pawelczyk, Klaus Broelemann, and Gjergji. Kasneci. On counterfactual explanations under predictive multiplicity. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 809–818. PMLR, 2020.

[Poyiadzi *et al.*, 2020] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[Rawal and Lakkaraju, 2020] Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. In *Advances in Neural Information Processing Systems*, 2020.

[Read and Marcus-Newhall, 1993] Stephen J. Read and Amy Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3):429–447, 1993.

[Russell, 2019] Chris Russell. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

[Singh *et al.*, 2023] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. Directive Explanations for Actionable Explainability in Machine Learning Applications. *ACM Transactions on Interactive Intelligent Systems*, pages 1–25, 2023.

[Thagard, 1989] Paul Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–467, 1989.

[Ustun *et al.*, 2019] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[Van Looveren and Klaise, 2021] Arnaud Van Looveren and Janis Klaise. Interpretable Counterfactual Explanations Guided by Prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665, 2021.

[Venkatasubramanian and Alfano, 2020] Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 284–293, 2020.

[Verma *et al.*, 2020] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. *arXiv preprint arXiv:2010.10596*, 2020.

[Wachter *et al.*, 2018] Sandra Wachter, Brent Mittelstadt, and Chris Russel. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and thhe GDPR. *Harvard Journal of Law & Technology*, 20(3):842–887, 2018.

[Wang *et al.*, 2023] Zijie J. Wang, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. GAM Coach: Towards Interactive and User-centered Algorithmic Recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[Warren *et al.*, 2023] Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 171–187, 2023.

[Yacoby *et al.*, 2022] Yaniv Yacoby, Ben Green, Christopher L. Griffin Jr., and Finale Doshi-Velez. "If it didn't happen, why would I change my decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pages 219–230, 2022.

[Yeh, 2016] I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.