Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on Human-Centred AI

Psychological Benefits and Costs of Diversifying Algorithmic Recourse

Tomu Tominaga[1] , **Naomi Yamashita**[2] , **Takeshi Kurashima**[1]
[1]NTT Human Informatics Laboratories, NTT, Inc.
[2]Graduate School of Informatics, Kyoto University
tomu.tominaga@ntt.com, naomiy@acm.org, takeshi.kurashima@ntt.com

Abstract

Algorithmic recourse provides counterfactual action plans that help people overturn unfavorable AI decisions. While diverse recourse sets may improve transparency and motivation, they may also impose cognitive load and negative emotions by increasing counterfactual reasoning demands. To examine this trade-off, we conducted a between-subjects controlled experiment (N=750) that manipulated recourse-set diversity and size, and evaluated these effects on psychological benefits and costs. Results show that diversification enhances psychological benefits (e.g., willingness to act) for small sets without incurring additional psychological costs, whereas for large sets, it makes cognitive load more salient. These findings suggest that naively diversifying recourse can burden decision subjects, underscoring the need for new diversification methods that incorporate human cognition and psychology to mitigate such costs.

1 Introduction

As AI models are increasingly being deployed in high-stakes domains, algorithmic recourse has become a central topic in explainable AI as a way to address the "right to explanation" [Wachter *et al.*, 2018]. Recourse takes the form of counterfactual explanations that suggest actionable changes allowing decision subjects to flip adverse decisions [Ustun *et al.*, 2019; Karimi *et al.*, 2021]. For example, a loan applicant denied by an AI system might be told, "If your annual income were $1,000 higher, you would be approved." Ideally, such recourse should be both understandable as an explanation and actionable as a behavioral plan [Karimi *et al.*, 2022].

To this end, a promising approach is to diversify recourse options by generating multiple counterfactuals that are mutually dissimilar in the feature space [Mothilal *et al.*, 2020]. Prior work suggests that presenting diverse recourse options can help people grasp complex concepts [Wang *et al.*, 2019; Spiro *et al.*, 1989], increase the chance that at least one plan is feasible under real-world constraints [Barocas *et al.*, 2020], and strengthen autonomy and motivation by letting decision subjects actively choose among alternatives [Ryan and Deci, 2000]. However, counterfactual reasoning often imposes cognitive demands and emotional burdens because it forces individuals to consider negative and positive outcomes simultaneously [Byrne, 2005; Byrne, 2007] or recall alternative pasts and personal shortcoming [Byrne, 2016; Tominaga *et al.*, 2025]. Although the trade-off between psychological benefits and costs is theoretically anticipated, it has rarely been examined empirically. Consequently, it remains unclear how recourse should be diversified in practice.

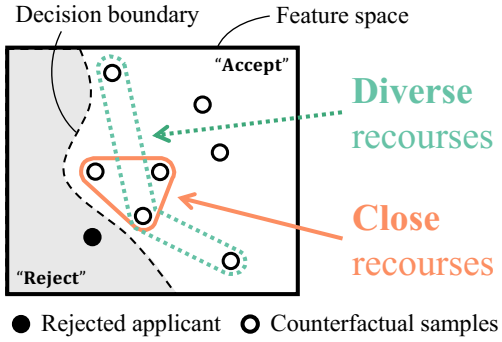

Figure 1: Conceptual difference in counterfactual sample selection between "Close" and "Diverse" conditions, illustrated with an example of three options. While the "Close" condition only focuses on proximity of counterfactuals, the "Diverse" condition focuses on both proxmity and mutual dissimilarity of counterfactuals.

To address this gap, we conducted a between-subjects experiment (N=750) in the context of automobile loan applications, manipulating both the number and diversity of recourse options. We generated sets of size 1, 3, or 7 using two methods: (1) selecting counterfactuals closest to the participant's data ("Close" condition), or (2) selecting counterfactuals close to the data but dissimilar from one another ("Diverse" condition). Figure 1 illustrates the conceptual difference between these two approaches. Participants evaluated the recourse set with regard to subjective reasonability, subjective actionability, willingness to act, and decision acceptance as indicators of benefits, and cognitive load and emotional burdens as measures of costs. This design allowed us to address the following research question (RQ): ***How do***

The supplementary materials are available at the following link: https://github.com/tomu-tominaga/PsyBenCosDivAR

recourse-set diversity and size affect decision subjects' psychological benefits and costs?

Addressing this RQ, we found contrasting patterns for benefits and costs. Diverse sets improved benefits from one to three options but then plateaued, while Close sets demonstrated stepwise gains from one to three to seven, matching Diverse sets only at seven. This trend was particularly pronounced in willingness to act. In contrast, cognitive load remained for Close sets beyond three, whereas it increased with set size for Diverse sets and became most salient at seven. Open-ended responses echoed this pattern; diverse sets helped participants find a meaningful plan at three options, but at seven they more often questioned the clarity of the AI's decision criteria. Taken together, these results suggest that diversification is promising for relatively small sets, but calls for safeguards when presenting larger sets.

This study makes three contributions: (1) we provide empirical evidence on how recourse diversity shapes participants' perceptions and experiences, bridging the gap between theoretical expectations and practice, (2) we identify the conditions under which the balance between benefits and costs is optimized, showing that diversification with relatively few options enhances motivation while minimizing psychological strain, and (3) based on these insights, we offer design implications to maximize the use of diversification, and discuss future directions for establishing such methods.

2 Related Work

To aid decision subjects who receive adverse AI judgments by clarifying its reasons and suggesting actionable steps, most research frames recourse generation as identifying the nearest counterfactual sample, an individual with a positive outcome whose profile is close to the target user [Wachter *et al.*, 2018]. Formally, given an N-dimensional feature space $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_N$, a binary classifier $f : \mathcal{X} \to \{+1, -1\}$, a distance function $\mathrm{dist} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, the task is to find the closest $\boldsymbol{c}^*$ in the positive region $\mathcal{A} = \{\boldsymbol{c} \in \mathcal{X} \mid f(\boldsymbol{c}) \neq f(\boldsymbol{x})\}$ for a target user $\boldsymbol{x} \in \mathcal{X}$ with $f(\boldsymbol{x}) = -1$:

$$\boldsymbol{c}^* \in \underset{\boldsymbol{c} \in \mathcal{A}}{\operatorname{argmin}}\, \mathrm{dist}(\boldsymbol{c}, \boldsymbol{x}), \tag{1}$$

and then construct the perturbation $\boldsymbol{\delta} = \boldsymbol{c}^* - \boldsymbol{x}$ as a recourse and provide it to the user. However, it is challenging to provide decision subjects with recourse both acceptable and actionable [Tominaga *et al.*, 2024], and researchers have proposed many techniques to tackle this problem (see extensive reviews [Karimi *et al.*, 2022; Verma *et al.*, 2024]).

Among them, one promising direction is to diversify recourse options [Laugel *et al.*, 2023]. This approach generates and presents multiple, varied recourse options so that decision subjects can identify an option that aligns with their circumstances and preferences [Wachter *et al.*, 2018; Barocas *et al.*, 2020]. Prior work proposes computational algorithms that force different recourse options to suggest different actions [Ustun *et al.*, 2019; Russell, 2019] or introduces the term that maximizes pairwise dissimilarity among counterfactuals in the feature space [Mothilal *et al.*, 2020].

Diversification is expected to benefit decision subjects in two ways: as explanations and as recommendations. As explanations, diverse recourse options may help decision subjects infer decision criteria because multifaceted explanations can improve understanding [Wang *et al.*, 2019; Spiro *et al.*, 1989]. Since better understanding is associated with stronger acceptance of unfavorable outcomes [Tominaga *et al.*, 2025], a diverse set may also make the explanation more persuasive. As recommendations, diversification can support both actionability and motivation. Specifically, a broader set of options increases the chance that at least one option is actionable under real-world constraints [Wachter *et al.*, 2018; Barocas *et al.*, 2020]. In addition, comparing alternatives and choosing a best fit can strengthen autonomy, a key source of motivation [Ryan and Deci, 2000]; therefore, offering diverse recourse options can enhance decision subjects' autonomy and thereby increase their willingness to act.

At the same time, diversification may impose cognitive and emotional costs. Counterfactual thinking can be more demanding than causal thinking [Byrne, 2005; Byrne, 2007] as it requires people to reason about dual possibilities (e.g., if X' instead of X, then Y' instead of Y) [McEleney and Byrne, 2006]. It can also elicit negative emotions because contrasting reality with counterfactual alternatives often evokes regret, guilt, or self-blame [Byrne, 2016]. In fact, recent work found that participants sometimes perceived recourse as blaming their past shortcomings [Tominaga *et al.*, 2025]. Since diversification exposes decision subjects to more varied counterfactuals, these psychological costs may intensify, especially as the number of recourse options increases.

Taken together, while diverse recourse sets are promising, they may also raise concerns. This trade-off has not yet been empirically examined, leaving open how much diversity should be introduced. Our study provides empirical evidence on this issue and identifies conditions for optimal diversity control in algorithmic recourse.

3 Method

To address our research question, we conducted a randomized between-subjects experiment manipulating the diversity and size of the set of recourse options, and elucidated the effects of these factors on psychological benefits and costs.

3.1 Experimental Setup

Hypothetical Scenario

Based on prior studies [Tominaga *et al.*, 2024; Tominaga *et al.*, 2025], we crafted a hypothetical scenario of a car loan screening because credit assessment is a central high-stakes domain in the research areas of machine learning [Karimi *et al.*, 2022; Verma *et al.*, 2024] and human-computer interaction [Wang *et al.*, 2023; Gemalmaz and Yin, 2022; Lyons *et al.*, 2022]. Specifically, participants engaged in the following scenario: *Participants were asked to imagine applying for a two-year auto loan equal to one-third of their annual income and submit their profile data to a financial institution that uses AI-based credit assessment. They then received a rejection notice with a recourse set and were instructed to evaluate its content.* In this scenario, the institution rejects loan applications when the requested amount is at least one-quarter of the applicant's annual income. Thus, our

#	Profile item (feature)	Con.	Typ.
1	Residence (e.g., Tokyo)	–	cat
2	Type of residence (e.g., rental)	–	cat
3	Education (e.g., high school)	↑	ord
4	Employment (e.g., private company)	–	cat
5	Job title (e.g., employee)	↑	ord
6	Years of service (e.g., 1-3 years)	↑	ord
7	Management career (e.g., 1-3 years)	↑	ord
8	Daily working hours (e.g., 0-2 hours)	–	ord
9	Daily remote working hours (e.g., 0-2 hours)	–	ord
10	Number of side jobs (e.g., 2 jobs)	–	ord
11	Job change experience (e.g., yes)	↑	ord
12	Overseas work experience (e.g., yes)	↑	ord
13	Study abroad experience (e.g., yes)	↑	ord
14	Best TOEIC† score (e.g., 600-695)	↑	ord
15	Facebook use (e.g., yes)	–	cat
16	LinkedIn use (e.g., yes)	–	cat

Table 1: Profile items used in the scenario. "Con." indicates the allowed direction of change in recourse (↑: increase-only; –: no direction constraint), which defines the feasible/mutable zone $\mathcal{Z}$ (Equation (1)). "Typ." denotes feature type: categorical (cat) or ordinal (ord). †TOEIC is a widely used English proficiency test in Japan.

experiment simulates a setting where all participants were rejected and would be approved if their profile corresponded to an annual income at least 4/3 times their current income. This rule is not disclosed to them.

The repayment period was fixed at two years, and the loan size was defined relative to each participant's income. This avoided variation in distance from participants to the decision boundary, which could affect their recourse evaluations.

Real Profile Data for Screening

Under this scenario, participants submitted their real profile data. The profile data items are shown in Table 1. Based on prior studies [Tominaga *et al.*, 2024; Tominaga *et al.*, 2025], we adopted 16 items: basic attributes (#1–#3), current employment (#4–#10), job experience and skills (#11–#14), and personal networks (#15–#16). These features commonly appear in loan studies [Yurrita *et al.*, 2023], credit evaluation datasets [Becker and Kohavi, 1996; Yeh, 2016; Hoffman *et al.*, 2013], and typical resumes [Brown and Campion, 1994; Cole *et al.*, 2007]. For the details of answer options and their feature values, see Supplementary Materials (SM), Table S1.

Immutable attributes such as gender were not included as they cannot be altered by individual effort [Kirfel and Liefgreen, 2021]. We also excluded participants' income from the recourse attributes. In our scenario, all participants were rejected for insufficient income. If income were included, all recourses would propose increasing it, sharply limiting diversity. To observe user reactions to diverse recourses, we did not use income as a profile item in recourses.

3.2 Factor Design and Manipulations

Factors

This study employed a two-factor factorial design manipulating recourse-set diversity and size. Recourse-set diversity captures the extent to which multiple recourse options are mutually distant in the feature space. To isolate the effect of diversity from the number of options, we independently varied set size, enabling a precise test of how diversity shapes participants' reactions. Each factor was designed as follows.

Diversity: Close or Diverse. To control recourse-set diversity, we constructed two types of recourse sets: (1) a *Close* set, formed by selecting counterfactual samples based solely on proximity to the decision subject [Wachter *et al.*, 2018], and (2) a *Diverse* set, formed by selecting samples based on proximity and pairwise dissimilarity among the selected samples [Mothilal *et al.*, 2020].

Size: 1, 3, or 7. We varied the number of recourse options as 1, 3, and 7. The single-option condition served as a reference condition for evaluating multi-option effects. We set the minimum set size to three. This is because, compared with only two options, three or more options enable comparisons across multiple pairs, facilitating the grouping of options [Tversky, 1977; Medin *et al.*, 1993] and thus helping individuals capture the spread and relational structure of options in the sample space, thereby enhancing the perception of diversity [Kahn and Wansink, 2004]. We set seven as an upper bound, expecting cognitive load to saturate near this range. Conventional work on short-term memory highlights limits on retaining and comparing "7±2" items [Miller, 1956], and later evidence shows that too many options can be ignored or yield random choices [Scheibehenne *et al.*, 2010].

Recourse Computation to Experimental Conditions

In this experiment, we randomly assign participants to one of the experimental conditions defined by two factors: recourse-set diversity $p := \text{Close}, \text{Diverse}$ and size $k := 1, 3, 7$. Hereafter, we denote each condition using the p–k notation (e.g., Diverse–3). Since diversity is undefined for a singleton set, we exclude the Diverse–1 condition, resulting in five conditions in total (i.e., $2 \times 3 - 1$).

To construct counterfactual samples for recourse sets, we used an existing dataset of 4,057 profile entries collected in prior work [Tominaga *et al.*, 2024]. We treated each record as a candidate profile $\boldsymbol{c} \in \mathcal{X}$ and restrict candidates to an admissible area $\mathcal{A} \cap \mathcal{Z}$. $\mathcal{A}$ denotes the set of distinct profiles that would be approved; in our loan scenario, this corresponds to profiles that satisfy the income-related approval condition (Section 3.1). $\mathcal{Z}$ encodes constraints on the direction of change for each feature. As shown in Table 1, we imposed these constraints to ensure that recourse suggestions remain plausible (e.g., "Education" is allowed to increase only).

Given a participant profile $\boldsymbol{x} \in \mathcal{X}$, we select a set of k counterfactual samples $C := \{\boldsymbol{c_1}, ..., \boldsymbol{c_k}\}$ by solving:

$$C^* = \underset{\substack{C \subset (\mathcal{A} \cap \mathcal{Z}) \\ |C|=k}}{\operatorname{argmin}} \frac{1}{k} \sum_{\boldsymbol{c} \in C} \operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}) - \lambda \cdot \frac{1}{\binom{k}{2}} \sum_{\substack{\boldsymbol{c_i}, \boldsymbol{c_j} \in C \\ i<j}} \operatorname{dist}(\boldsymbol{c_i}, \boldsymbol{c_j}), \tag{2}$$

where the first term captures proximity (i.e., average distance to $\boldsymbol{x}$) and the second term measures diversity (i.e., average pairwise distance). We set $\lambda = 1$ for Diverse and $\lambda = 0$ for Close, so the diversity term is considered only for Diverse. We then compute the recourse set $\Delta = \{\boldsymbol{c} - \boldsymbol{x} \mid \boldsymbol{c} \in C^*\}$.

Inspired by prior studies [Mothilal *et al.*, 2020; Wachter *et al.*, 2018], dist (Equation (2)) is defined as a feature-wise ℓ_1 distance between given samples $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{X}$:

$$\mathrm{dist}(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{|\mathcal{I}_{\mathrm{cat}}|} \sum_{i \in \mathcal{I}_{\mathrm{cat}}} \mathbb{1}[\boldsymbol{u}_i \neq \boldsymbol{v}_i] + \frac{1}{|\mathcal{I}_{\mathrm{ord}}|} \sum_{i \in \mathcal{I}_{\mathrm{ord}}} \frac{|\boldsymbol{u}_i - \boldsymbol{v}_i|}{\mathrm{MAD}_i}. \tag{3}$$

Here, $\mathbb{1}$ is the indicator function (1 if true, 0 otherwise). $\mathcal{I}_{\mathrm{cat}}$ and $\mathcal{I}_{\mathrm{ord}}$ are the sets of categorical and ordinal features, respectively; $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ denote the option values of feature i. The first term averages mismatches over categorical features, while the second term averages absolute differences over ordinal features, normalized by the feature-wise median absolute deviation MAD_i.

To solve Equation (2), we used a greedy approximation (see SM, Section A for pseudocode) because our goal is stimulus generation with controlled diversity rather than exact optimization. Importantly, this procedure produced a clear separation in perceived diversity in our manipulation check (Section 3.4), validating its adequacy for the user study.

3.3 Measured Outcomes

As described in Section 2, recourse-set diversification possibly brings various psychological benefits and costs to decision subjects. To capture them, this study used a questionnaire survey (see SM, Section B for the full instructions).

To improve the validity of our measurements, we included a comprehension task prior to the measurements (see SM, Section B.4). In this task, participants were instructed to identify the profile items that the recourse options cited as reasons for rejection and recommended actions for future approval. Their responses inconsistent with the presented recourse set were excluded from the analyses. This task screened out inattentive or low-effort responses, and ensured that subsequent ratings were based on an accurate understanding of the recourse. Psychological benefits and costs were measured after the comprehension task as follows.

Psychological Benefits

Assuming that decision subjects who are provided with a recourse set undergo a sequential experience of judging whether the suggested plan is a reasonable explanation [Tominaga *et al.*, 2024], assessing whether it is feasible to carry out [Kirfel and Liefgreen, 2021; Wang *et al.*, 2023], forming motivation to act [Tominaga *et al.*, 2024], and integrating these impressions into their ultimate acceptance of the AI decision [Tominaga *et al.*, 2025], we captured psychological benefits with the following four dimensions:

Subjective reasonability: the extent to which participants found the recourse option to be a reasonable explanation for the rejection.

Subjective actionability: the extent to which participants found the recourse option easy or difficult to carry out.

Willingness to act: the extent to which participants were willing to carry out the recourse option.

Decision acceptance: the extent to which participants accept the AI decision outcome of their loan application.

All items were rated on a 7-point scale. For subjective reasonability, subjective actionability, and willingness to act, we additionally asked participants to provide an open-ended rationale for each rating. We did not collect rationales for decision acceptance because it captures participants' overall acceptance of the decision outcome as a downstream response to the presented recourse evaluations (see SM, Section C for associations between decision acceptance and the other items).

For participants in multiple-option conditions, we used a selection-and-rating process for the three recourse-evaluation measures: for each measure, participants first selected the option they considered best on that dimension, and then rated the selected option and provided its rationale. This procedure was not applied in the single-option condition.

Psychological Costs

For cognitive load, this study used NASA-TLX [Hart and Staveland, 1988], which assesses subjective workload on experimental tasks with six dimensions, given its widespread use in HCI studies [Kosch *et al.*, 2023]. We adopted it to observe participants' cognitive load on the comprehension task. Among the six dimensions, we used the following three:

Mental demand: the extent to which mental and perceptual activity is required for the comprehension task.

Effort: the extent to which participants have to work hard to accomplish the comprehension task.

Frustration: the extent to which participants feel irritated and stressed to accomplish the comprehension task.

These are scaled in 5-point increments from 0 to 100. Given the context of the comprehension task with minimal physical activity, no strict time pressure, and no explicit emphasis on performance, we do not discuss the remaining dimensions—Physical Demand, Temporal Demand, and Performance (see SM, Section D for full results).

For emotional burdens, we observed:

Negative emotional experience: the number of emotions participants experience in reviewing the recourse options among regret, shame, guilt, self-blame, disappointment, disgust, dissatisfaction, and discrimination,

considering that counterfactual thinking can evoke these emotions as it highlights alternative pasts [Byrne, 2016], personal shortcomings [Tominaga *et al.*, 2024], or insufficient achievement [Tominaga *et al.*, 2025].

3.4 Procedure

This experiment was conducted from February to March 2025 in Japan. Figure 2 describes the overall procedure.

Participants. Participants were recruited via the Japanese online survey company. They were first informed of the study's purpose and content, and consent was obtained. To enhance ecological validity, we conducted the screening survey to select those who (1) work at a private company or public institution, (2) consider a car purchase, (3) hold no loans, and (4) earn less than 10 million JPY annually. In total, 1062 participated in the experiment. All were Japanese: 300 females, 762 males; 49.2 years old on average. They received compensation equivalent to 600 JPY for the participation.

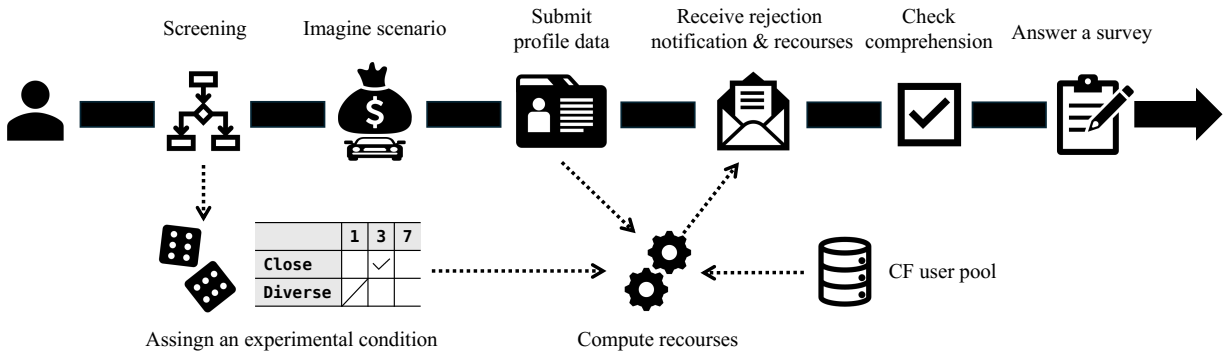

Figure 2: Overview of the experimental procedure.

	$k=1$	$k=3$	$k=7$
$p=$ Close	139	152	168
$p=$ Diverse	–	126	165

Table 2: Number of participants across the experimental conditions of the recourse-set diversity p and size k.

Recourse provision under the scenario. After the screening, we randomly assigned participants to one of the five experimental conditions. They were instructed to imagine the scenario and submit their real profile data (Section 3.1). Based on the participants' experimental condition and profile data, we computed the recourse set for each participant (Section 3.2). Participants were then shown a rejection notice along with the recourse set as an AI-generated plan for future approval (see SM, Figure S1 for an example).

Comprehension check. We administered the comprehension task (Section 3.3); participants' responses inconsistent with the presented recourse set were judged to have insufficient understanding and were excluded from subsequent analyses. As a result, 750 of the 1,067 participants were retained for analysis. Table 2 shows the distributions of participants across the experimental conditions.

Outcome measurement. The questionnaire survey (Section 3.3) first asked about cognitive load. Placing it immediately after the comprehension task allowed us to more precisely capture the cognitive load required for understanding the recourse. Participants then completed the questionnaires on psychological benefits and emotional burdens in sequence.

Manipulation check. To confirm that our manipulations effectively altered perceptions of diversity, we asked participants with the multiple-option conditions how diverse the presented recourse set on a 7-point scale, and confirmed that our manipulation was successful using a two-way ANOVA; perceived diversity was significantly greater for Diverse than Close sets and for 7 than 3 options (diversity: p=0.042; size: p=0.002; see SM, Table S16 for full statistics).

3.5 Analysis

For each outcome of psychological benefits and costs, we conducted a series of two-way ANOVAs with recourse-set diversity and size as factors. When the interaction was significant, we performed Tukey's HSD for post-hoc pairwise comparisons. As diversity is undefined for a singleton set, the single-option condition was not included in the two-way ANOVAs. Instead, we conducted Dunnett's multiple-comparison test with Close–1 as the reference, comparing it against all other conditions to quantify the effect of multiple recourse options relative to a single option.

To interpret the results qualitatively, we coded participants' open-ended responses. One author developed the codebook iteratively, and an external collaborator independently coded a stratified random 10% subsample per outcome. The coders achieved substantial agreement [Landis and Koch, 1977]: Cohen's $\kappa \geq 0.61$. The remaining responses were coded by the author and summarized into overarching themes.

4 Results and Findings

Figure 3 shows how recourse-set diversity and size affect psychological benefits and costs. Overall, while both benefits and costs tended to increase as the set size k grew, the growth patterns differed between Close and Diverse. Benefits increased up to k=3 and then plateaued in Diverse, whereas they rose more gradually with k in Close and converged to Diverse at k=7 (Figure 3 (A)–(D)). In contrast, cognitive load increased up to k=3 and then plateaued in Close, whereas they continued to grow with k in Diverse, reaching a level comparable to or higher than Close at k=7 (Figure 3 (E)–(G)). Emotional burdens were more mitigated in Diverse than Close (Figure 3 (H)).

Below, we describe three findings that unpack these patterns. For transparency, SM reports the full quantitative outputs (Section D, Figure S3, Tables S3–S13) and the complete qualitative results (Section E, Tables S14 for the codebook and S15 for the code distributions). We refer to rationale codes of each outcome measure using specific prefixes and IDs (e.g., SR-1, SA-1, WA-1). In this section, we focus on the key comparisons supporting Findings 1–3.

4.1 Finding 1: Benefit of Diversification for Willingness to Act Diminishes for Larger Sets

As shown in Figure 3 (C), recourse-set diversity had a size-dependent effect on willingness to act. A two-way ANOVA

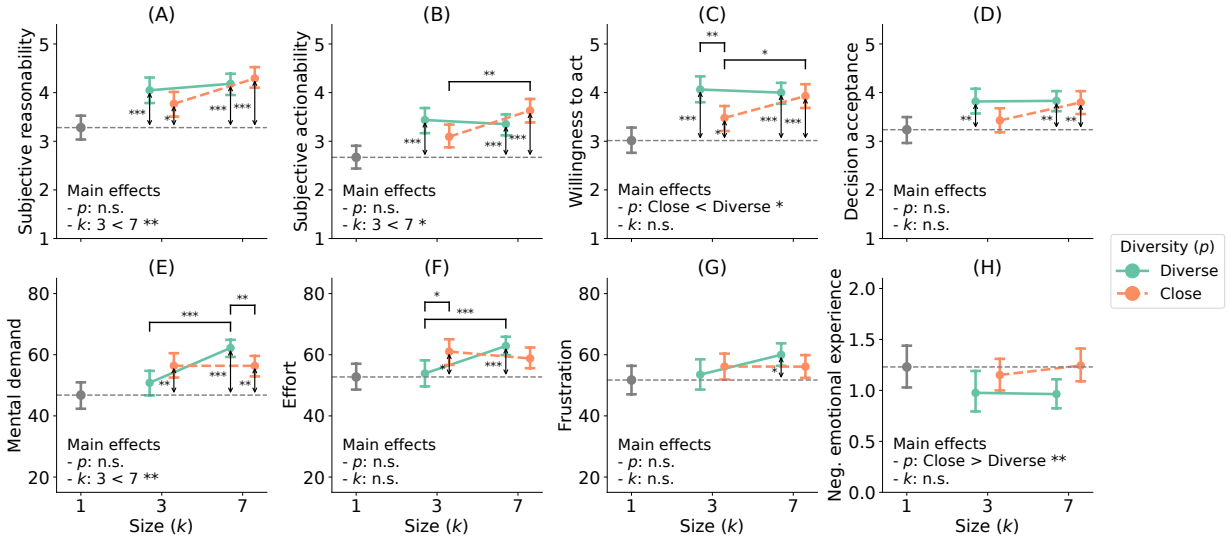

Figure 3: Psychological benefits (top row: (A)–(D)) and psychological costs (bottom row: (E)–(H)). We report (i) two-way ANOVAs with recourse-set diversity p and size k, followed by Tukey's HSD post-hoc tests, and (ii) Dunnett's one-to-many comparisons using Close–1 as the reference. All p-values are adjusted (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$). Main effects from the ANOVAs are reported in the lower-left of each panel. Tukey's HSD results are indicated by umbrella symbols, and Dunnett's comparisons are indicated by double-arrow symbols.

revealed a significant interaction effect of diversity and size ($F(1,607)=4.005$, $p=0.046$, $\eta_p^2=0.007$). Tukey's HSD post-hoc tests showed that Diverse outperformed Close only at $k=3$ ($p=0.002$), not at $k=7$. Consistently with this pattern, the qualitative analysis found that Diverse–3 more frequently mentioned motives and values (WA-6; 11.1% vs. 4.5%) and less often described the recourse as unnecessary (WA-7; 0.9% vs. 7.6%) than Close–3.

Taken together, diversification can yield its benefits by helping participants to find a motivating and necessary plan at a moderate set size ($k=3$), whereas its advantage diminishes as the set becomes larger ($k=7$).

4.2 Finding 2: Cognitive Load of Diversification Becomes Pronounced for Larger Sets

As shown in Figure 3 (E), mental demand exhibited a significant interaction ($F(1,607)=10.121$, $p=0.002$, $\eta_p^2=0.016$). Tukey's post-hoc comparisons indicated that Diverse exceeded Close at $k=7$ ($p<0.01$), suggesting that the cognitive cost of diversification becomes salient for larger sets. The similar directional patterns were observed for effort and frustration: Diverse<Close at $k=3$ and Diverse>Close at $k=7$ (Figure 3 (F), (G)). The qualitative rationales also showed that participants' references to lack of clarity about the decision criteria were less prevalent for Diverse than Close at $k=3$ (SR-5; 5.4% vs. 8.8%), whereas they were slightly more prevalent for Diverse at $k=7$ (SR-5; 7.8% vs. 6.9%).

To understand this, we observed inconsistency across recourse options, assuming that contradictory suggestions consume cognitive resources [Kivikangas *et al.*, 2025]. We here defined *sign conflict* as the presence of different options in a set changing the same feature in opposite directions (e.g., an option increases the number of side jobs and another decreases it in a same set). As shown in Table 3, Diverse often included sign conflict as set size increased. From a descriptive post-hoc analysis, we also found that adding an explanatory variable of sign conflict to the OLS model significantly improved the model fit for mental demand beyond the diversity and size (see SM, Table S17).

	Close–3	Diverse–3	Close–7	Diverse–7
Sign conflict	16.4%	11.1%	29.2%	35.8%
Common change	58.6%	35.7%	31.0%	27.9%

Table 3: Proportions of recourse sets with sign conflict (i.e., different options within a set change a same feature in opposite directions) or common change (i.e., all options within a set change a same feature) in each condition, as observed in exploratory analyses.

These results and observations suggest that diversification of larger sets ($k=7$) often induce within-set inconsistency, possibly imposing more cognitive load.

4.3 Finding 3: Emotional Burden Is Unlikely to Worsen under Diversification

As shown in Figure 3 (H), negative emotional experience was reduced under diversification regardless of set size ($F(1,607)=7.949$, $p=0.005$, $\eta_p^2=0.013$). This pattern runs counter to a naive expectation that increasing the number or diversity of options would amplify emotional burden (Section 2), and it suggests that emotional burdens behave differently from cognitive costs (Finding 2).

To interpret this, we focused on redundancy among recourse options, inspired by the idea that receiving the same advice repeatedly is unpleasant [Cacioppo and Petty, 1979; Kocielnik and Hsieh, 2017]. As its proxy, we explored *common change*, defined as the presence of a feature that is changed in a same direction by all options within a set. As

shown in Table 3, Diverse tended to exhibit fewer common changes than Close, especially at k=3. However, a descriptive post-hoc analysis showed that common change did not relate to negative emotional experience (see SM, Table S17).

To sum up, diversification was unlikely to exacerbate emotional burden; the drivers of this pattern warrant further investigation, including redundancy.

5 Discussion

As indicated by our findings, diversification can increase willingness to act for smaller sets, while cognitive load likely become salient for larger sets. This size-dependent pattern warns failure modes in generating large diverse sets and motivates design strategies to address them. This section describes its details, and then outlines limitations and future work.

5.1 Implications

The size-dependent trade-off between benefits and costs rests on two principles of diversification. First, as set size grows, the relative advantage of diversification shrinks. This is because, even when different policies are used (e.g., Close or Diverse), each chooses increasingly overlapping counterfactual samples. This convergence is likely as promising counterfactuals are scarce in practice [Keane and Smyth, 2020]. This principle is consistent with our result that the diversification benefits diminished for larger sets (Section 4.1).

Second, diversification has a structural property that prioritizes counterfactuals altering different features, and once these are exhausted, then chooses samples altering same features in opposite directions. This is optimal strategy to minimize the objective function in Equation (2). Consequently, inconsistent suggestions (e.g., sign conflict) likely emerge in large diverse recourse sets. Given that decision making under conflicting aims increases cognitive effort [Kivikangas *et al.*, 2025], sign conflict possibly contributed to the increased mental demand (Section 4.2). Notably, if the first principle was only considered, differences in cognitive load between the diversity conditions would shrink as set size grows. However, we observed that, rather than converging, cognitive load became more salient under diversification for larger sets.

These interpretations suggest that diversification may raise cognitive load when set size is ignored, calling for remedies to both the scarcity of promising samples and inconsistencies across recourse options. This perspective lends empirical support to prior technical directions, including mechanisms for securing sufficiently many promising samples in model construction [Kanamori *et al.*, 2024] or securing coherency among recourse options [Rasouli and Yu, 2024].

Close sets increased benefits with set size, but plateaued cognitive load (Figure 3), suggesting that multiple, similar but distinct suggestions may increase perceived persuasiveness. Since sets were constructed from unique profiles (Section 3.2), no sets presented duplicate options. Instead, repeating identical recourse options would likely attenuate the effect due to repetition discomfort [Cacioppo and Petty, 1979].

5.2 Recommendations

Our findings inform that the policy of recourse diversification algorithms in AI systems should consider the recourse-set size and properties. For relatively small sets, diversification is recommended because it can increase willingness to act while keeping cognitive and emotional burdens manageable. For relatively large sets, by contrast, it is desirable to suppress within-set inconsistency and determine the set size adaptively. Following this policy, we recommend extending the proximity–diversity objective by introducing set size and within-set inconsistency as either hard constraints or penalty terms, where inconsistency is captured by sign conflicts across recourse options.

5.3 Limitations and Future Directions

Our study adopted set size k=1, 3, 7 based on cognitive theory to create qualitatively distinct conditions as noted in Section 3.2, but we did not test intermediate sizes such as k=5; thus, we cannot precisely localize the set size at which the benefit-cost balance turns. Importantly, our contribution does not lie in the specific numbers per se (e.g., 3 or 7), but in elucidating the qualitative pattern that the benefits of diversification can plateau beyond a certain size while the costs continue to grow with k. Accordingly, k=1, 3, 7 should not be interpreted as generalizable "magic numbers," as the turning point may vary with the domain and dataset, and ultimately with the availability of promising counterfactual samples. To identify this turning point in practice, future work should calibrate it within the target task. Since our results suggest that costs increase with set size, such efforts should prioritize inspecting the plateau point of benefits.

We used an automobile-loan setting as the experimental task; therefore, there remains room for future work to examine whether our findings can be applied to different decision-making domains. In addition, our sample was skewed toward men and drawn from a single cultural context. As decision-making styles [Hofstede, 2011] and trust in AI [Maslej *et al.*, 2025] may vary across cultures, further studies should test whether similar results hold among decision subjects with greater demographic and cultural diversity.

Translating our empirical findings into concrete technical solutions remains future work. As noted in Section 5.2, one promising direction is to incorporate set properties (e.g., size or sign conflict) into recourse generation objective. These directions would help establish generation and presentation techniques that automatically tune the benefit-cost balance.

6 Conclusion

Through a randomized controlled between-subjects experiment (N=750) complemented with quantitative and qualitative analyses, this paper provides empirical evidence on the size-dependent trade-off between psychological benefits and costs of diversifying algorithmic recourse. Specifically, diversification increased willingness to act for smaller sets, whereas it made cognitive load more pronounced for larger sets. Our post-hoc analysis suggests a key pitfall: diversifying larger sets often induces inconsistencies across options, which may amplify cognitive load. To address this, we highlighted the potential of diagnostic properties for recourse verification and optimization. We hope these foundational findings and discussions will pave the way for human-centered AI systems integrating human cognition and psychology.

Ethical Statement

This study involving human subjects was reviewed and approved by the external ethics review board, the Institutional Review Board of Public Health Research Foundation (approval ID: PHRF-IRB 25A0004), and all procedures were conducted in accordance with the guidelines.

References

[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. ACM, 1 2020.

[Becker and Kohavi, 1996] Barry Becker and Ronny Kohavi. Adult, 1996.

[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*, 79:897–908, 12 1994.

[Byrne, 2005] Ruth M. J. Byrne. The rational imagination: How people create alternatives to reality. *The Rational Imagination*, 6 2005.

[Byrne, 2007] Ruth M.J. Byrne. Précis of the rational imagination: How people create alternatives to reality. *Behavioral and Brain Sciences*, 30, 12 2007.

[Byrne, 2016] Ruth M.J. Byrne. Counterfactual thought. *Annual Review of Psychology*, 67:135–157, 1 2016.

[Cacioppo and Petty, 1979] John T. Cacioppo and Richard E. Petty. Effects of message repetition and position on cognitive response, recall, and persuasion. *Journal of Personality and Social Psychology*, 37:97–109, 1 1979.

[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*, 56:319–343, 4 2007.

[Gemalmaz and Yin, 2022] Meric Altug Gemalmaz and Ming Yin. Understanding decision subjects' fairness perceptions and retention in repeated interactions with ai-based decision systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 295–306. ACM, 7 2022.

[Hart and Staveland, 1988] Sandra G. Hart and Lowell E. Staveland. *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*, volume 52, pages 139–183. North-Holland, 1 1988.

[Hoffman *et al.*, 2013] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. Trust in automation. *IEEE Intelligent Systems*, 28:84–88, 1 2013.

[Hofstede, 2011] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2:8, 12 2011.

[Kahn and Wansink, 2004] Barbara E. Kahn and Brian Wansink. The influence of assortment structure on perceived variety and consumption quantities. *Journal of Consumer Research*, 30:519–533, 3 2004.

[Kanamori *et al.*, 2024] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Learning decision trees and forests with algorithmic recourse. In *Forty-first International Conference on Machine Learning*, 2024.

[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. ACM, 3 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, 2022.

[Keane and Smyth, 2020] Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *Case-Based Reasoning Research and Development (ICCBR)*, pages 163–178, 2020.

[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.

[Kivikangas *et al.*, 2025] J. Matias Kivikangas, Eeva Vilkkumaa, Julian Blank, Ville Harjunen, Pekka Malo, Kalyanmoy Deb, Niklas J. Ravaja, and Jyrki Wallenius. Effects of many conflicting objectives on decision-makers' cognitive burden and decision consistency. *European Journal of Operational Research*, 322:182–197, 4 2025.

[Kocielnik and Hsieh, 2017] Rafal Kocielnik and Gary Hsieh. Send me a different message: Utilizing cognitive space to create engaging message triggers. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2193–2207. ACM, 2 2017.

[Kosch *et al.*, 2023] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys*, 55:1–39, 12 2023.

[Landis and Koch, 1977] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159, 3 1977.

[Laugel *et al.*, 2023] Thibault Laugel, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Achieving diversity in counterfactual explanations: a review and discussion. In *2023 ACM Conference on Fairness Accountability and Transparency*, pages 1859–1869. ACM, 6 2023.

[Lyons *et al.*, 2022] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. What's the appeal? perceptions of review processes for algorithmic decisions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15. ACM, 4 2022.

[Maslej *et al.*, 2025] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. The ai index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, 2025.

[McEleney and Byrne, 2006] Alice McEleney and Ruth M. J. Byrne. Spontaneous counterfactual thoughts and causal explanations. *Thinking & Reasoning*, 12:235–255, 5 2006.

[Medin *et al.*, 1993] Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. Respects for similarity. *Psychological Review*, 100:254–278, 4 1993.

[Miller, 1956] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 3 1956.

[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. ACM, 1 2020.

[Rasouli and Yu, 2024] Peyman Rasouli and Ingrid Chieh Yu. Care: coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, 17:13–38, 1 2024.

[Russell, 2019] Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28. ACM, 1 2019.

[Ryan and Deci, 2000] Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55:68–78, 2000.

[Scheibehenne *et al.*, 2010] Benjamin Scheibehenne, Rainer Greifeneder, and Peter M. Todd. Can there ever be too many options? a meta-analytic review of choice overload. *Journal of Consumer Research*, 37:409–425, 10 2010.

[Spiro *et al.*, 1989] Rand J. Spiro, Paul J. Feltovich, Richard L. Coulson, and Daniel K. Anderson. *Multiple analogies for complex concepts: antidotes for analogy-induced misconception in advanced knowledge acquisition*, pages 498–531. Cambridge University Press, 7 1989.

[Tominaga *et al.*, 2024] Tomu Tominaga, Naomi Yamashita, and Takeshi Kurashima. Reassessing evaluation functions in algorithmic recourse: An empirical study from a human-centered perspective. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pages 7913–7921. International Joint Conferences on Artificial Intelligence Organization, 8 2024.

[Tominaga *et al.*, 2025] Tomu Tominaga, Naomi Yamashita, and Takeshi Kurashima. The role of initial acceptance attitudes toward ai decisions in algorithmic recourse. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20. ACM, 4 2025.

[Tversky, 1977] Amos Tversky. Features of similarity. *Psychological Review*, 84:327–352, 7 1977.

[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. ACM, 1 2019.

[Verma *et al.*, 2024] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56:1–42, 12 2024.

[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*, 31:842–887, 2018.

[Wang *et al.*, 2019] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15. ACM, 5 2019.

[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*. Association for Computing Machinery, 2023.

[Yeh, 2016] I-Cheng Yeh. Dafeault of credit card clients, 2016.

[Yurrita *et al.*, 2023] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 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*, pages 1–21. ACM, 4 2023.