アルゴリズミックリコースの受容可能性と実行可能性に関する暗黙的仮定の実証的検証

冨永 登夢[1,a)]　山下 直美[2]　倉島 健[1]

概要：AI システムから否定的な判定を受けた個人が肯定的な判定結果を得るための行動プラン –リコース– を提案するアルゴリズミックリコースの究極的な目標は，対象者にとって受け入れやすく（受容可能性が高く）実行しやすい（実行可能性の高い）リコースの生成である．これまでのリコース生成技術は（1）個人の現状と目標状態の間の距離を最小化するリコースは受容可能性と実行可能性が高い，（2）全ての個人はリコースに対して 1 つの共通する距離関数を持つ，という 2 つの仮定を前提としてきた．ところが，これらの仮定について実証的に検証された結果はほとんど報告されていない．そこで本研究は，362 名を対象とした調査によりこれらの仮定の妥当性を検証した．その結果，第 1 の仮定に反して，受容可能性はリコース距離と独立であり，実行可能性はリコース距離が最小の時最大値を示すがそれ以外では一定であることがわかった．さらに，第 2 の仮定も支持されず，否定的な事象に対して敏感な性格を持つ被験者はリコースに対して受容可能性と実行可能性を高く評価する傾向にあることが確認された．これらの結果からリコース生成研究の根本的前提を再検討し，ユーザ中心型リコース生成技術の設計指針について論じた．

1. はじめに

アルゴリズミックリコースは，AI システムから望ましくない判定結果を受けた個人を救済するため，その判定結果を覆すための行動プランを個人に提供する XAI 技術である [21]．具体的には，個人が望ましい出力を AI システムから得るために必要な変更を示す “リコース” と呼ばれる反実仮想的な提案を行う．例えば，ローン審査に落ちた人に対して “もしあなたの年収が 100 万円増加すれば，ローン申請は承認されるでしょう” といった情報を提示する．アルゴリズミックリコースの最終的な目標は，対象者がそれを受け入れて行動を起こせるようなリコースの提案である．ここで，判定結果の説明としてのリコースの受け入れやすさを受容可能性，リコースで提案されている行動の実行しやすさを実行可能性と定義する．

受容可能性と実行可能性の高いリコースを生成するために，リコース生成研究は対象者との乖離が最も小さい反実仮想サンプルを特定するタスクを解く．乖離を定量化する距離関数の例として，対象サンプルから反実仮想サンプルに変更される特徴量の数を示す *Sparsity* や対象サンプルと反実仮想サンプルの距離を表現する *Proximity* がある [49]．このタスクは最適化問題として定式化され [21], [49], [51]，これまで数多くの技術的解法が急速に提案されている [21], [49].

このような技術的発展の一方で，これらの基盤となる根幹的な 2 つの仮定はいまだに実証されていない．その 1 つ目は，対象サンプルと反実仮想サンプルの距離を最小化するリコースは受容可能性と実行可能性が高い，という仮定である [46], [51]．この仮定は，人々はより単純な説明を好む [36], [42] という Miller の指摘 [30] に基づいている．ところが，これは “直感に基づくアプローチへの過度な依存” [28] であり，Sparsity や Proximity などの距離関数の最小化が受容可能性と実行可能性を保証するか否かは実証されていない．例えば，ローン申請承認のために「学歴」と「勤続年数」と「副業数」の向上を提案するリコース（Sparsity=3）と「居住地」の変更を提案するリコース（Sparsity=1）がある時，対象者は後者のリコースを受け入れて実行すると従来研究は仮定するが，このような前提が真かどうかは不明瞭である．頑健な科学的進歩のためには，実際の被験者を対象とした調査によってこの仮定を検証し，心理学的に裏付ける必要がある．

第 2 の仮定は，全ての個人がある 1 つの共通する距離関数を持つ，とするものである．これは明らかに過度な単純化であり，対象者が提案されたリコースを好むかどうかは個人によって異なる [3], [25], [48], [53]．例えば，否定的な

[1] 日本電信電話 (株) NTT 人間情報研究所
NTT Human Informatics Laboratories, NTT Corp.
[2] 日本電信電話 (株) NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corp.
[a)] tomu.tominaga@ntt.com

出来事に対して繊細な人は AI システムから否定されるような自身の現状を避けたいと思うことでリコースに高い受容可能性や実行可能性を示しうる．つまり，仮に「勤務時間」「学歴」「勤続年数」「副業数」「居住地」など変更項目が多数なリコースは一般的には好まれないとしても，上記のような対象者は自身の現況を改善したいと考えリコースを受け入れるかもしれない．しかし，今日までに受容可能性および実行可能性と個人の性格の関係を系統的に調べた結果は報告されていない．この検証は，受容可能性と実行可能性の高い個別化されたリコースを生成する技術の確立において重要な課題であると我々は考えている．

そこで本論文は以下の研究課題を解くことで，上記の未検証仮定に関する実証的な知見を得ることを目的とする．

RQ1. リコースの受容可能性及び実行可能性は距離関数とどのような関係にあるか？それはなぜか？

RQ2. 人格特性はリコースの受容可能性及び実行可能性と距離関数の関係にどのような影響を与えるか？

我々は，自動車ローン申請の場面におけるリコースの受容可能性と実行可能性を評価するオンライン調査を 362 名を対象に実施した．また，否定的な出来事への敏感さに関連する，回避性人格障害 [54]，神経症傾向 [57]，悲観主義 [58] を本実験の人格特性として対象とした．

結果として（1）第 1 の仮定に反して，**受容可能性はリコース距離と独立の関係にある**こと，**実行可能性はリコース距離が最小の時に最大値を示すがそれ以外では一定である**こと，（2）第 2 の仮定に反して，やや距離の遠いリコース（例えば Proximity=2）に対して**神経症傾向や回避性人格障害の傾向が強い被験者は高い実行可能性を，悲観主義傾向の強い被験者は高い受容可能性を示す**ことが分かった．

本研究の貢献は，（1）被験者実験を通じてリコース生成研究における重要な仮定について検証し，潜在的な欠陥について実証的な知見をもたらしたこと，（2）ユーザの人格特性がリコースに対する受容可能性と実行可能性に及ぼす影響を明らかにしたこと，（3）得られた知見からユーザ中心型リコース生成技術の設計指針と倫理的配慮事項を示したことにある．

2. 関連研究

2.1 アルゴリズミックリコース

リコース生成研究の目的は，AI システムから望ましくない判定を受けた個人がその決定を理解し改善するための方法を確立することにある [21]．これを実現するため，アルゴリズミックリコースは反実仮想的な提案を示すことで，対象となる個人が AI システムの否定的な判定結果を受け入れ [17], [51]，その提案にしたがって判定結果を覆すためのその後の行動を起こすように促す [20], [22]．

個人にとって受け入れやすくまた実行しやすいリコースは簡潔で簡単な変更を示すはずだと考えられている [46], [51]．この着想は，人はなるべく言及する原因の数が少ないようなシンプルな説明を好む [36], [42] と述べた Miller のレビュー論文 [30] に基づいている．この考え方のもと，リコース生成研究は以下の最適化問題（式 1）を解く [21], [49], [51]：N 次元の特徴量空間 $X = X_1 \times \cdots \times X_N$ を入力とする予測モデル（例えば二値分類モデル）を $h : X \to \{-1, +1\}$ として，入力サンプル $\boldsymbol{x} \in X$ に対する反実仮想サンプル $\boldsymbol{x'} \in X$ を特定しリコースを生成する．なお，$S(X) \subseteq X$ はドメイン知識 [20] などの任意の条件によって定められる X の部分空間である．

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

このタスクにおいて $d(\cdot, \cdot) : X \times X \to \mathbb{R}_{\geq 0}$ は入力サンプル $\boldsymbol{x}$ と反実仮想サンプル $\boldsymbol{x'}$ の乖離を定量化する距離関数である．最も原始的な距離関数として L_0 ノルム（特徴量の変更数）と L_1 ノルム（マンハッタン距離）がある [49]．本稿では前者を Sparsity，後者を Proximity と呼ぶ．

これまでのリコース生成研究の多くが技術的性能の向上に焦点を当てており [21], [49]，特徴量間の因果関係を制約項に導入する [22], [23]，一貫した多様な説明を生成する [31], [37]，変更行動の順序を算出する [19], [32] といった様々な解法を上記タスクに対して提案してきた．これらの手法では，最適化問題における距離関数の最小化により対象とする個人にとって受容可能性と実行可能性の高いリコースを生成できると仮定されているが，これを支持する実証的知見は乏しい．本研究ではこれらの仮定を検証する調査を行う．

2.2 アルゴリズミックリコースに対する人間中心の評価

アルゴリズミックリコースや反実仮想説明の研究領域における人間中心の評価の不足が指摘されており [18], [35]，リコースの受容可能性と実行可能性を直接的に評価した論文はほとんど存在しない．この評価の難しさの 1 つとして，高いリスクを伴う意思決定を下す場面で対象者が AI システムから望ましくない判定を受けるという状況の再現が実験的に困難であるという点が挙げられる．

この課題に取り組んだ上で反実仮想的提案の主観評価を試みた数少ない研究 [27], [53] を紹介する．Wang らは，LendingClub データセット [1] で学習されたモデルを用いてローン審査を行うインタラクティブなリコース生成システムを開発し，被験者に当データセットに含まれるユーザになったつもりで AI システムから否定的な結果を受け取り，その後生成されたリコースを評価するように依頼するシナリオ実験を行った結果，被験者は操作可能で小さい変更を好むことを確認した [53]．しかし，これは自分以外の誰かのふりをする代理的な意思決定であり，本人が自身の状況を踏まえて下す意思決定ではない．

対照的に，Kuhl ら [27] は実験環境としてゲームを利用

することで，被験者自身の（操作）データに基づく反実仮想的な提案を被験者に評価させる実験を行った．この実験では，被験者はエイリアンが飼育されている架空の動物園の飼育員としてプレイするゲームに参加する．この実験用ゲームでは，効率的な餌やりの戦略を見つけてエイリアンの数を増やすというミッションがプレイヤーに与えられており，そのための参考情報としてプレイデータに基づいた反実仮想的な提案が定期的にゲーム側からプレイヤーに対して提供される．Kuhl らは被験者がミッション達成のためにどのような提案を参考にするかを調査した [27]．ここで注目すべきは，プレイデータを元に構成された反実仮想的提案の受容可能性をユーザ本人が評価した点である [27]．ただしゲーム環境に依拠するため，実世界の文脈は考慮されていない．

本研究は（1）実世界の文脈を踏まえた [53]，かつ（2）対象者自身のデータに基づいた [27] 反実仮想的提案を対象とする．これにより従来研究よりも精緻にリコースの受容可能性と実行可能性を評価する．

2.3 リコースの評価に影響を与える人格特性

受容可能性や実行可能性は主観的な概念である [3], [24], [48] ため，対象者の心理的な特徴と関連する可能性が高い．アルゴリズミックリコースは AI システムによる望ましくない判定結果に直面した個人を対象にするという性質から，本研究では否定的な事象に対する対象者の認識，感情，そして反応に注目する．

例えば，悲観主義傾向にある人は悪い結果を予想する傾向にある [38], [58] ため，AI システムによる不利な判定を受けて自身の現状をより否定的に捉えてしまうかもしれない．同様に，神経症傾向（情緒不安定性）が高い人も否定的な情報に敏感である [15], [43] ことから，望ましくない判定結果に対して不安や心配といった感情になりやすい可能性がある．さらに，回避性人格障害を抱える人は他者からの批判や非難を避けようとする傾向にある [2], [54] ため，自身に対する AI システムの否定的な出力結果に強い抵抗感を示すと予想される．

このような性格を持つ個人が AI から否定的な判定を受けた後にリコースをどのように評価するかは明らかにされていない．これらの性格は引っ込み思案 [10], [11] や自己効力感の低さ [45] と関連するため，彼らは AI システムの否定によって動機を失うかもしれない．一方で，そのような性格を持つ人は他者に対して従順であるという性質を持つ [9], [13], [29] ため，AI から否定された自身の現状を変えるための具体的なプランであるリコースに対して強い納得感や実行意欲を示す可能性も考えられる．

本稿では，悲観主義，神経症傾向，回避性人格障害の 3 つの人格特性をネガティブセンシティビティと呼び，受容可能性および実行可能性との関係性について調査する．

3. 実験

3.1 実験参加者

アスマーク社のモニターから募った 362 名を実験参加者とした（平均年齢は 47.3 歳，うち 147 名（40.6%）が女性）．参加条件は（1）一般企業の会社員もしくは公的機関の職員である（正規非正規問わず），（2）自動車の購入を検討している，（3）現在ローンを組んでいない，（4）年収が 1 千万円未満である，の全てを満たすこととした．第 1 条件は所得のある個人を対象とすることを目的としている．第 2，第 3 条件は，自動車を購入する動機もしくは経済的な余裕がないために，いかなるリコースに対しても実行可能性を低く評価してしまう被験者を除外するために定めた．第 4 条件は，本実験の被験者に対する反実仮想サンプルを事前に収集したユーザデータプールの中から選択できるように設定した（3.3.1 節参照）．参加者は事前に実験内容に関する同意確認書を著者らに提出した．実験終了後，参加者は 100 円相当の謝礼を受け取った．

3.2 実験設計とデータ

3.2.1 シナリオ設定

本実験では，被験者にはある金融機関にて自動車ローンを申請する場面を想定してもらう．この場面を選定した理由は，リコース研究のドメインとして金融領域 [20], [46], [53] や信用審査 [26], [53] が頻繁に採用されるからである．また，融資されたお金で購入する対象物のうち（例えば，教育ローン，自動車ローン，住宅ローンなど），自動車は多くの人にとって馴染みがあり，比較的高額であり，主に対象者自身によって所有され利用されるため，自動車ローン申請のシナリオは被験者にとって，その申請状況を想像しやすい，高リスクの意思決定であると認識しやすい，実世界の文脈を考慮しやすい場面であると考えられる．

具体的には以下のような状況を再現する．

> 実験参加者は，自身の年収の 1/3 相当の自動車購入のため 2 年のローンを組もうと考えている．彼らは，ローンの融資を受けるためにある金融機関を訪れ，そこで審査のためのプロファイルデータを提出するように求められた．その審査手続きには AI システムが導入されており，申請者がローン融資を受けられるか否かを金融機関に記録されている膨大な顧客データに基づいて当 AI システムが判定している．彼らはプロファイルデータを提出して審査を受けたが，申請結果は不承認となった．そこで彼らは，本申請が承認されなかった理由と承認されるために必要な行動を知るために，AI システムが彼らのデータに基づいて作成したリコースをいくつか見せてもらうことにした．

ここで我々は実験的に 2 つの前提を置いている．第 1 に，この金融機関は申請者の年収の 1/4 以上を要求するローン融資申請を全て不承認とするというルールを持つこととした．そのため本実験では，全参加者の申請が不承認となることに注意されたい．この審査基準は実験全体を通じて被験者に知らされない．第 2 に，被験者に提示するリコースを生成するための反実仮想サンプルのデータプールとして 3683 名分のユーザプロファイルデータを事前に収集し（3.3.1 節参照），当シナリオにおける “膨大な顧客データ” に相当するデータベースとして用意した．

本シナリオにおけるローンの金額は絶対値ではなく年収に対する相対値である．また，返済期間も 2 年に固定されている．この実験設計は，被験者間におけるローン返済負担の大きさを統制し，返済負担が実行可能性の評価に影響しないように定められている．これらが事前に定められていない場合，例えば年収 500 万円の個人が融資額 400 万円で返済期間 1 年のローン申請を行うという非現実的な状況が起こる可能性があり，被験者がリコース距離とは無関係にいかなるリコースに対しても実行可能性を低く評価してしまうと危惧される．これらの実験設計は，このような影響の除外を目的としている．

3.2.2 属性情報，購買意欲，ローン状況

我々は事前に実験参加者の性別，年齢，職業，年収，自動車購買意欲，ローン借入状況をアンケートで取得した．職業，年収，自動車購買意欲，ローン借入状況の情報は実験参加条件（3.1 節参照）の判定に用いた．

3.2.3 人格特性

本実験で対象とする人格特性は，回避性人格障害 [54]，神経症傾向 [57]，および悲観主義 [58] の 3 つである．回避性人格障害については，Williams と Benjamin が開発した 7 設問から構成される質問紙の邦訳版 [54] を用いた．各設問は 3 段階の回答選択肢を持つため，回避性人格障害の傾向の強さは 7〜21 点となる（被験者平均：12.22±3.89）．

神経症傾向については，Gosling らが回答者の負担を軽減するために開発した 10 設問で Big Five に含まれる 5 因子のスコアを測定する TIPI [15] の邦訳版 TIPI-J [57] を用いた．神経症傾向を測定するための 2 設問はそれぞれ 7 段階の回答選択肢を持つため，神経症傾向の強さは 2〜14 点となる（被験者平均：7.56±2.39）．

悲観主義については，外山が開発した楽観性/悲観性尺度を測定する質問紙 [58] を用いて計測した．これに含まれる悲観性尺度を計測する 10 設問を利用した．各設問は 4 段階の回答選択肢を持つため，悲観主義の強さは 10〜40 点となる（被験者平均：7.56±2.39）．

3.2.4 プロファイルデータ

本研究の実験においてリコースに含まれるプロファイルデータを表 1 に示す．属性情報（#1-3），現職状況（#4-10），スキルと経験（#11-14），人脈（#15,16）に関連する項目が含まれる．我々は，与信審査のための機械学習データセットの特徴量 [1], [4], [16], [56] や典型的なレジュメに含まれる情報 [6], [8] を参考にこれらの項目を定めた．

ここで，不可変特徴量に相当する情報は採用していない．不可変特徴量とは，年齢，性別，出生地，人種など，本人によって操作・変更することのできない情報を指す [26]．このような特徴量の変更を提案するリコースはしばしばユーザの理解を促すが，常に実行不可能となる [21]．本実験では受容可能性と実行可能性の双方を評価するため，これらの特徴量はリコースから除外した．

また，年収情報もリコースには含めていない．本実験では，被験者のローン申請は申請額に対する年収の不足により不承認となるというシナリオが用いられる．そのため，年収情報をリコースに含めた場合，変更する特徴量の数（Sparsity）が 1 つであるリコースは必ず年収増加の変更を提案することになる．これにより，リコース評価の対象が Sparsity ではなく年収そのものになってしまう問題が生じる．これを回避するために我々は年収情報をプロファイルデータから除外することで，様々な特徴量の変更が Sparsity=1 のリコースで提案されるように対処した．

3.2.5 リコースの評価

被験者がリコースの受容可能性と実行可能性をどのように評価するかを確認するため，我々は以下の 2 つの設問を用意した：“AI システムによって提示されたプランは自身のローン申請が不承認となった理由の説明として納得感がありますか？”（受容可能性），“AI システムによって提示されたプランを実行すればあなたのローン申請は承認されますが，実行しようと思いますか？”（実行可能性）．いずれの設問に対しても 7 段階尺度の回答選択肢を設けた（1. 全く思わない – 7. 強く思う）．各設問の後に，被験者はリコースに対する評価の理由を “なぜそのように評価したのですか？” という自由記述回答式の設問で回答した．

さらに，リコースで提案されている各特徴量の変更の不可変性についても尋ねた．前述の通り我々は一般に不可変特徴量 [26] と想定される項目を事前にリコースから除外しているが，表 1 に含まれる項目が個別の被験者にとっては不可変である可能性がある．そのため我々は，リコースで提示されたそれぞれの特徴量の変更が不可能であるかどうかを，はい/いいえの二択で被験者に尋ねた．この設問で「はい」と回答された特徴量が含まれるリコースは分析から除外した．

3.3 評価用リコースの構成

3.3.1 反実仮想サンプル用のユーザデータプール

入力サンプル $\boldsymbol{x}$（式 1）となる被験者のプロファイルデータに対応する反実仮想サンプル $\boldsymbol{x'}$（式 1）を特定してリコースを構成するためには，事前に反実仮想サンプルの候補となるプロファイルデータの集合を用意する必要があ

表 1 リコース用プロファイルデータ（x_i と x'_i は入力サンプルと反実仮想サンプルの要素 i）

#	項目（特徴量）	選択肢（オプション）	制約条件	尺度
1	居住地	1. 東京 / 2. 東京以外	$x'_1 \geqq x_1$	名義
2	居住形態	1. 持ち家 / 2. 賃貸・寮	$x'_2 \geqq x_2$	名義
3	最終学歴	1. 高校卒 / 2. 短大/専門卒 / 3. 大学卒 / 4. 院卒（修士） / 5. 院卒（博士）	$x'_3 \geq x_3$	順序
4	勤務先	1. 一般企業 / 2. 公的機関	$x'_4 \geqq x_4$	名義
5	職位	1. 一般社員 / 2. 主任 / 3. 係長 / 4. 課長 / 5. 次長 / 6. 部長 / 7. 本部長・事業部長 / 8. 常務取締役 / 9. 専務取締役 / 10. 代表取締役	$x'_5 \geq x_5$	順序
6	勤続年数（年）	1. 0-1 / 2. 1-3 / 3. 3-5 / 4. 5-10 / 5. 10-20 / 6. 20-	$x'_6 \geq x_6$	順序
7	マネジメント歴（年）	1. なし / 2. 0-1 / 3. 1-3 / 4. 3-5 / 5. 5-10 / 6. 10-20 / 7. 20-	$x'_7 \geq x_7$	順序
8	勤務時間（時間/日）	1. 0-2 / 2. 2-4 / 3. 4-6 / 4. 6-8 / 5. 8-10 / 6. 10-12 / 7. 12-	$x'_8 \geqq x_8$	順序
9	在宅勤務時間（時間/日）	1. 0-2 / 2. 2-4 / 3. 4-6 / 4. 6-8 / 5. 8-10 / 6. 10-12 / 7. 12-	$x'_9 \geqq x_9$	順序
10	副業数	1. なし / 2. 1 / 3. 2 / 4. 3 / 5. 4 / 6. 5-	$x'_{10} \leqq x_{10}$	順序
11	転職歴	1. なし / 2. あり	$x'_{11} \geq x_{11}$	順序
12	海外勤務歴	1. なし / 2. あり	$x'_{12} \geq x_{12}$	順序
13	海外留学歴	1. なし / 2. あり	$x'_{13} \geq x_{13}$	順序
14	TOEIC 最高得点	1. なし / 2. 10-400 / 3. 400-495 / 4. 500-595 / 5. 600-695 / 6. 700-795 / 7. 800-895 / 8. 900-990	$x'_{14} \geq x_{14}$	順序
15	Facebook 利用	1. 未登録・友達なし / 2. 登録済・友達あり	$x'_{15} \geqq x_{15}$	名義
16	LinkedIn 利用	1. 未登録・友達なし / 2. 登録済・友達あり	$x'_{16} \leqq x_{16}$	名義

る．そこで我々は，2023 年 6 月 12 日から 29 日の期間にアスマーク社のモニターから一般企業の会社員もしくは公的機関の職員を募り，対象者の年収と表 1 のプロファイルデータを収集した．収集終了後，全ての回答者に対して 30 円相当の謝礼を支払った．なお，年収について尋ねる際には「回答したくない」という回答選択肢を設けた．最終的に 4057 件の回答があり，そのうち年収について回答されたプロファイルデータは 3683 件となった．以降において，反実仮想サンプル用のユーザデータプールとは，この 3683 件のプロファイルデータを指す．

本実験のシナリオにおける金融機関の審査基準（3.2.1 節参照）の性質上，年収額の大きい被験者に対して評価用のリコースを用意できない可能性がある．例えば，年収 1000 万円の被験者が参加した場合には，実験設計上データプール内の年収 1332 万円以上のプロファイルデータが反実仮想サンプルの候補となるが，このようなプロファイルデータは 1.47%（54/3683）しか含まれていない．このような事態を回避するため，我々はシナリオ実験において実験参加者の年収上限を設けた（3.1 節参照）．

3.3.2 リコース構成のルール

入力サンプル $\boldsymbol{x}$ に対して以下の 2 つの条件を満たす反実仮想サンプル $\boldsymbol{x'}$ を選択する：(1) $\boldsymbol{x'}$ の年収は $\boldsymbol{x}$ の 4/3 倍以上である，(2) $\boldsymbol{x'}$ と $\boldsymbol{x}$ は表 1 の制約条件にしたがう．第 1 条件はシナリオ実験における金融機関（AI システム）の審査基準に対応している．この条件を満たせばローン申請額は年収の 1/4 未満となり申請は承認されるため，これを満たす反実仮想サンプルがリコースの対象となる．

第 2 条件は，起きえない特徴量の変更を除外する．例えば，最終学歴，職位，勤続年数，マネジメント歴といった項目は下げる（減らす）ことができない．これは，転職歴，海外勤務歴，海外留学歴，TOEIC 最高得点といった経験やスキルについても同様である．そのためこれらの特徴量値は $\boldsymbol{x}$ よりも $\boldsymbol{x'}$ の方が大きいという制約条件を課す．これら以外の項目については制約条件を設けていない．

3.3.3 リコースの距離尺度と選択基準

これらのルールによって抽出されたリコースの中から，以下の手順にしたがって 5 つの反実仮想サンプルを選択する：(1) Sparsity が最小のサンプルを 1 つ選ぶ，(2) Proximity が最小のサンプルを 1 つ選ぶ，(3) ランダムに 3 つ選ぶ．ここで，入力サンプル $\boldsymbol{x}$ とその反実仮想サンプル $\boldsymbol{x'}$ に対して，摂動ベクトル $\boldsymbol{\delta}$ を

$$\delta_i = \begin{cases} \mathbb{I}[x'_i \neq x_i] & (\text{特徴量 } i \text{ が名義尺度}) \\ |x'_i - x_i|/M_i & (\text{特徴量 } i \text{ が順序尺度}) \end{cases} \tag{2}$$

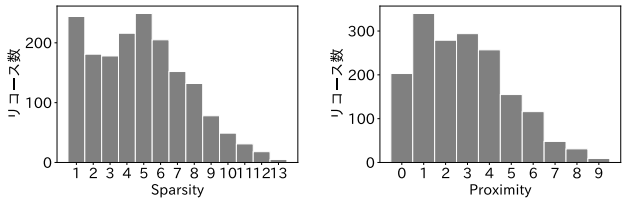
と定める．ここで，M_i は特徴量 i の最大変更範囲を指しており，特徴量 i の摂動量 δ_i は最小値が 0・最大値が 1 となるように正規化される．これに対し，リコースの距離として Sparsity と Proximity を以下のように算出する．

$$\text{Sparsity} = \sum_i \mathbb{I}[\delta_i \neq 0], \quad \text{Proximity} = \sum_i \delta_i. \tag{3}$$

選択された評価用リコースの距離の分布を図 1 に示す．

3.4 実験手順

本シナリオ実験は 2023 年 7 月 7 日から 8 月 10 日にオンライン上で実施された．実験参加者はまず実験用の基本情報（3.2.2 節参照）と人格特性（3.2.3 節参照）について回答した．その後，本実験のシナリオを想定するように依頼され，プロファイルデータ（3.2.4 節参照）を提出した．提



(a) Sparsity (b) Proximity

図 1 評価用リコースの距離尺度の分布

出後，審査のため 2 週間待機するように指示される．その間に我々は被験者ごとにリコースを構成し（3.3 節参照），リコース評価用のアンケートを作成し，被験者に審査結果として送付した．被験者は評価用アンケートの通知を受け取った際，自身のローン申請が不承認だったことを想定するよう指示される．その後，被験者は評価用アンケートに記載されている 5 つのリコースを評価した（3.2.5 節参照）．

3.5 分析

我々は合計で 1810 件のリコースに対する評価値を収集した．そのうち特徴量変更が不可変と評価された 72 件のリコースを除外し，残りの 1738 件を分析に用いた．

RQ1 の定量的評価として，我々は受容可能性/実行可能性と Sparsity/Proximity 間の依存関係を確認するために相関分析を行なった．また，リコースを距離尺度に応じて群分けし，各群間における受容可能性/実行可能性の差をウェルチの t 検定を用いて検証した．これは多重比較に相当するため，ボンフェローニ補正によって有意水準を調整した．また，評価理由に関する自由記述回答（3.2.5 節参照）から被験者のリコースに対する認識を確認し，定量的分析結果との関連性を調べた．

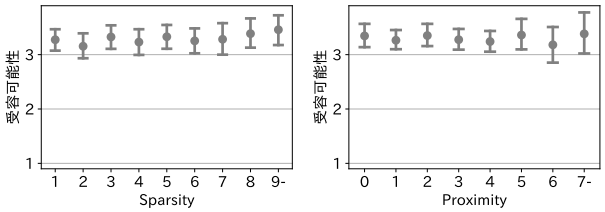
RQ2 に答えるため，我々はまず人格特性ごとにそのスコアの平均値を基準として被験者を 2 群（高群，低群）に分けた．次に，RQ1 と同様にリコースを距離尺度に応じて群分けした．その後，ある特定距離のリコース群において，受容可能性/実行可能性が 2 つの人格特性群間でどの程度異なるかをウェルチの t 検定で比較した．なお，RQ1 と同様に有意水準をボンフェローに補正によって調整した．この分析により，リコース距離を統制した上で人格特性の高さが受容可能性/実行可能性に与える影響を評価した．

4. 結果

4.1 リコース距離と受容可能性/実行可能性（RQ1）

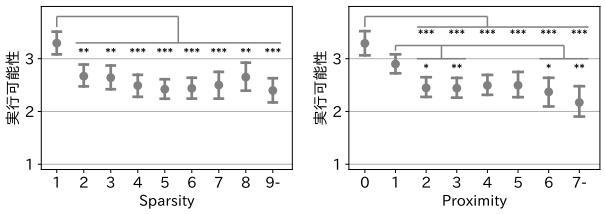
4.1.1 受容可能性

図 2 にリコース距離と受容可能性の関係を示す．相関分析の結果，受容可能性はいずれの距離尺度とも無相関であった（Sparsity：$r = 0.036$, $p = 0.139$；Proximity：$r = 0.001$, $p = 0.977$）．また，ウェルチの t 検定（多重比較）の結果，いずれのリコース群間においても受容可能性の高さに差は確認されなかった．つまり，距離尺度が小さいほど受容可能性は高いという従来研究の仮定は支持されない結果となった．



(a) 受容可能性 vs Sparsity (b) 受容可能性 vs Proximity

図 2 受容可能性と距離尺度の関係．エラーバーは 95%信頼区間．



(a) 実行可能性 vs Sparsity (b) 実行可能性 vs Proximity

図 3 実行可能性と距離尺度の関係．エラーバーは 95%信頼区間．調整済み有意確率：*...$p < 0.05$, **...$p < 0.01$, ***...$p < 0.001$.

距離尺度の小さいリコースの受容可能性を低く評価した被験者の自由記述回答を確認すると，“それだけ？って感じ”（P446），“現状と比較して特に変わり映えがしない”（P706）といったように，彼らは自身のプロファイルが反実仮想サンプルとほとんど同様なのにも関わらず申請結果が不承認となったことに懐疑的であった．一方で，距離尺度の大きいリコースの受容可能性を高く評価した被験者は，“歴然とした差があるから”（P732），“理由としてしっかりしてるので仕方ないと思う”（P739）といったように，変更量の多いリコースを見て自身の至らなさを実感し，納得して受け入れていたことが分かった．

4.1.2 実行可能性

図 3 にリコース距離と実行可能性の関係を示す．相関分析の結果，実行可能性は距離尺度と弱い負の相関関係にあることが分かった（Sparsity：$r = -0.128$, $p < 0.001$：Proximity：$r = -0.159$, $p < 0.001$）．また，リコース群間を比較した結果，Sparsity=1 の群はそれ以外のどの群よりも高い実行可能性を示した（図 3(a)）．Proximity=0 の群は Proximity≥2 のどの群よりも，Proximity=1 の群は Proximity=2,3,6,7 の群よりも高い実行可能性を示すことが分かった（図 3(b)）．これら以外のリコース群間に実行可能性の差は認められなかった．つまり，Sparsity/Proximity の低いリコースの実行可能性は高いという従来研究の仮定は概ね支持されるが，リコース距離が一定以上の領域（Sparsity≥2，Proximity≥2）では実行可能性はリコース距離とは無関係に一定の値を示すという結果となった．

距離尺度の小さいリコースに対する実行可能性の評価理由として，“今すぐ取り組めそうだからです”（P674），“勤

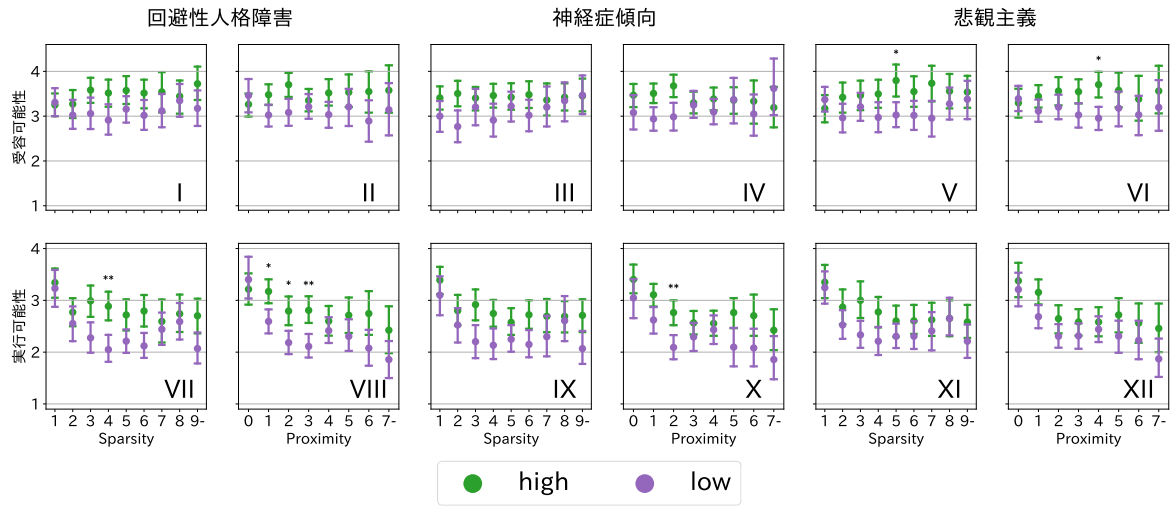


図 4 人格特性によるリコース距離と受容可能性/実行可能性の関係への影響．エラーバーは95%信頼区間．調整済み有意確率：*...$p < 0.05$, **...$p < 0.01$, ***...$p < 0.001$.

務時間だけだったら、すぐにでもやっていきたい”（P718），“簡単に変更できる内容だったので”（P294）といったように，即時的にもしくは簡単に実行できる提案であると認識していた記述が多く見られた．反対に，距離尺度の大きいリコースに対しては，“やることが多すぎて時間やお金がかかるため”（P401），“クリアする項目が多すぎる”（P167），“色々変えてまでローン申請したいとは思わない”（P468）といった報告が見られ，時間的・経済的コストの大きさ，変更量の多さ，費用対効果の悪さなどの負担の重さに言及する評価理由が述べられていた．

4.2 人格特性の影響（RQ2）

図 4 にリコース距離と受容可能性/実行可能性の関係が人格特性によってどのように異なるかを示す．

回避性人格障害は，受容可能性に対して影響しないが，実行可能性に対して有意な影響を与えた．具体的には，回避性人格障害のスコアが高い被験者は，そうでない被験者に比べて，Sparsity=4（図 4VII：$t = 4.252$，$p = 0.002$）と Proximity=1,2,3（図 4VIII：$t = 3.332$，$p = 0.049$；$t = 3.370$，$p = 0.045$；$t = 3.990$，$p = 0.004$）のリコース群に対して高い実行可能性を示した．回避性人格障害と同様に，神経症傾向も実行可能性に対してのみ有意な影響を与えることが分かった．神経症傾向のスコアの高い被験者は，そうでない被験者に比べて，Proximity=2 のリコース群に対して高い実行可能性を示した（図 4X：$t = 3.823$，$p = 0.008$）．悲観主義は，上記の 2 つの人格特性とは異なり，実行可能性ではなく受容可能性に対して影響した．悲観主義のスコアの高い被験者は，そうでない被験者に比べて，Sparsity=5（図 4V：$t = 3.415$，$p = 0.039$）と Proximity=4（図 4VI：$t = 3.742$，$p = 0.012$）のリコース群に対して高い受容可能性を示すことが分かった．

5. 考察

5.1 リコース生成問題の再検討

5.1.1 受容可能性のあるリコース生成に向けて

本実験の受容可能性に関する定量的および定性的分析から示された通り，距離関数の最小化は受容可能性を保証しない．ゆえに，現状の距離関数で受容可能性を評価するという研究方針は信頼できるアプローチとは言えない．

受容可能性を保証する有効な戦略を考察するため，我々は小さな変更を提案するリコースに対して受容可能性を高く評価した理由と低く評価した理由の違いを調べた．ここでは，学歴の変更を提案する Sparsity=1 のリコースを対象とした．その結果，低評価の理由として被験者は “給与と関係ないところなので納得できない”（P293），“学歴がローンの審査に必要だと思わない”（P468）と報告していた．一方，高評価の被験者は “学歴はステータスになると普段から考えているから”（P287），“学歴が年収に関わるから”（P401）といった理由を報告していた．

これらの内容は，提案された変更が判定結果と関連するか否か，もしくはその関連性がユーザの事前知識と一貫するか否かが受容可能性に影響を与えることを示唆している．これを踏まえると，リコースの受容可能性を保証する有効なアプローチは 2 つあると考えられる．1 つ目は，特徴量と判定結果の因果関係を捉えて制約項に導入するものである．これに関する技術的解法は，Karimi らの一連の研究 [21], [22], [23] ですでに議論され始めている．

2 つ目は，ユーザとのインタラクションを通じてユーザが判定結果と関連しないと考える特徴量もしくはその変更を特定し，リコース生成時にそれらを除外するというアプローチである．前述のように特徴量や判定結果に関わる因果関係の検出は有望だが，一般的に精緻な因果関係の特定

は極めて難しい課題であり，仮に特定できたとしてもそれがユーザの知識と衝突するとリコースの受容可能性は低下してしまう．そのため，先行研究に示されるような直接的な対話によってユーザの嗜好を捉えるインタラクティブシステム [41], [52], [55] によって，ユーザの事前知識と調和するリコースを生成する仕組みの確立が重要である．

5.1.2 実行可能性のあるリコース生成に向けて

我々は，実行可能なリコースを生成するために距離関数の値域に制約を課すことを推奨する．これは，距離関数の最小化による実行可能性の最大化は，距離関数が特定の閾値以下の値を出力する場合に限られており，それ以外では距離関数と実行可能性は独立であることが示唆されたからである．つまり，特定の閾値以上の領域において，距離関数は実行可能性を捉えることができないのである．そのため，距離関数の値域に上限を設けることで，強制的に実行可能性を保証できる．このアプローチはいくつかの既存研究で試験的に用いられており [34], [47]，本研究はその妥当性を実証的に示すものと位置付けられる．

5.2 ユーザ中心型リコース生成技術の設計指針

受容可能で実行可能なリコースを個別的に生成するためには，ユーザの事前知識に適応する機能と人格特性を活用する機能の 2 つを備えた計算機システムが有望である．

5.2.1 ユーザの事前知識への適応

前者の機能の有効性は，個々のユーザに対応してリコースを生成するインタラクティブシステムを開発した研究 [53] の結果から期待される．具体的にはこの先行研究では，特徴量変更の難易度やリコースの直感性に関するユーザフィードバックをリコース生成プロセスに組み込むことで，ユーザが自身にとって好ましいリコースを探索できたこと，自身の嗜好をシステムに反映できるインタラクティブデザインを高く評価したことが確認されている [53].

このようなインタラクティブな設計 [41], [52], [53], [55] によって，変更難易度や直感性に加えて特徴量と判定結果との関連性についてユーザから情報を得ることができれば，対象者の事前知識と衝突しないリコースを生成できると期待される．ただし，これが実際に受容可能性と実行可能性の保証に寄与するか否かは後続の検証実験が必要である．

5.2.2 ユーザの人格特性の活用

様々なメディア情報を用いて人格特性を検出するための研究 [5], [14] は多く，性格を捉えることでユーザにとって調和的な個別化されたインタラクションを実現できる [12], [33], [50] と期待される．本研究の結果もまた，ユーザの人格特性データを活用することでユーザにとって好ましいリコースを適応的に生成できる可能性を示唆する．

具体的には，距離関数の値域の上限を人格特性に基づいて定めれば，個々人の性格に応じて実行可能なリコースを探索する範囲を設定できる．例えば，実行可能性が 2.5 以上となるリコースの生成を目的としてユーザの回避性人格障害の回答値を利用する場面を想定する．ここで，図 4VIII から，回避性人格障害の傾向が強いユーザには Proximity=4 を，弱いユーザには Proximity=1 を距離関数の上限とする制約を課すことで，上記の目的を達成するリコースの探索範囲を性格特性に応じて設定できる（前者のユーザ群には Proximity≤4，後者のユーザ群には Proximity≤1）．一般に，ユーザは複数のリコースを提示される場合が多い [31], [37] ため，このような探索範囲を適応的に個別化できるアプローチは有望である．

ただし，ユーザの人格特性は事前に質問紙で測定される必要がある．第三者が利用可能なデータで人格特性を推定する手法 [14] も存在するが，我々は質問紙による取得を推奨する．本実験では，回避性人格障害を 7 問 [54]，神経症傾向を 2 問 [57]，悲観主義を 10 問 [58] で計測した．人格特性は一般に時間の経過に対して安定的であるため，事前取得した情報をユーザデータとして登録しておくことで上記のアプローチは実現可能である．ただし，これらの質問数がユーザにとって負担でないか，負担を抑えるために質問数を減らしても時間安定性は確保されるかについては将来研究で検証される必要がある．

5.3 アルゴリズミックリコースにおける倫理的配慮事項

RQ2 の結果は，ネガティブセンシティビティを有する個人は AI システムの提案にしたがってしまいやすいという潜在的な危険性を示唆する．この解釈は以下の 4 点に基づいている．第 1 に，本実験参加者は AI システムから否定的な判定結果を受けるという好ましくない状況に強制的に置かれる．第 2 に，回避性人格障害や神経症傾向を抱える人は批判やストレスに敏感 [15], [54] で他者の提案に倣いやすい [13], [29] ことが知られている．第 3 に，ネガティブセンシティビティはやや距離尺度の大きいリコース，すなわち少なくない変更量を要求するリコースに対する評価に影響を与えている．第 4 に，実行可能性が高いことと受容可能性が高いことは必ずしも対応しない．例えば，回避性人格障害は Sparsity=4 のリコースに対する実行可能性を高める効果を持つ（図 4VII）が，受容可能性を高める効果はない（図 4I）．

これらを踏まえると，回避性人格障害や神経症傾向を持つ個人は AI システムによって否定された望ましくない現状を避けることを優先しており，そのような状況下で AI システムから行動プランであるリコースを提案されると，それが仮にコストのかかる受け入れづらいものだったとしても，高い実行意欲を示す傾向にあると考えられる．この結果は本人にとって負担となる行動を納得感のないままに従事させるシステムの設計につながる危険性を暗示する．

難しい状況下にあるユーザを救済しようとする知的情報システムに関する研究は対象ユーザがこのようなリスク

に晒されていないか注意深く観察する必要がある．ここには，求職者に対する情報推薦 [7]，差別的な経験をした個人に対するケア [40], [44]，ハラスメントへの対処 [39] といった課題解決を図る研究が含まれる．我々は，このような介入支援がネガティブセンシティビティを抱える対象者を過剰に動機づけていないか観察することを推奨する．

5.4 本研究の制約

本研究では金融機関における信用審査に関わる場面を実験的にシナリオとして採用した．そのため，医療診断や人事雇用といった異なるドメインにおける本研究の知見の再現性は今後検証されなければならない．加えて，本シナリオにおけるローン返済期間（2 年）は一般的なものと比較すると短く設定されている．我々はローン経験のない被験者を対象としたためこの影響はほとんどないと予想されるが，そうでない被験者を対象とする場合にはより長期的な期間が望ましい．また，本分析に用いた距離関数は非常に原始的であるため，より複雑な距離関数を採用した際に同様の結果が確認されるかどうかは今後調べなければならない．さらに，本実験は日本人を対象に日本人用に開発された質問紙調査を用いたため，異なる文化言語圏で同様の実験を実施し，知見の一貫性を確認することが重要である．

6. 結論

本研究はアルゴリズミックリコース生成技術の研究領域における 2 つの根幹的な未検証仮定について，362 名を対象とした被験者実験により実証的知見を示した．予想に反し，リコース距離関数は受容可能性を捉えておらず，また，一部の領域以外では実行可能性を保証しないという結果となった．さらに，ネガティブセンシティビティを抱える被験者は，AI に否定された現状から脱却しようとして，やや変更量の多いリコースに高い実行可能性を示すことが示唆された．これらの結果に基づいて，ユーザ中心型リコース生成技術の設計指針や倫理的配慮事項について論じた．

本研究における実証的知見はこれまでのリコース生成技術の基本的前提を再検討するだけでなく，新たな最適化問題の設計，個別化されたリコース生成システムの新機能の実装，リコースに対する心理的な反応のさらなる系統的調査といった将来的な研究の方向性を示すものである．我々はこの研究が今後の XAI 研究の科学的な進歩と技術的な発展の礎になることを切に願う．

参考文献

[1] : LendingClub.
[2] American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition*, American Psychiatric Publishing, Inc. (1994).
[3] Barocas, S., Selbst, A. D. and Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons, *FAT**, pp. 80–89 (2020).
[4] Becker, B. and Kohavi, R.: Adult, UCI Machine Learning Repository (1996).
[5] Berkovsky, S., Taib, R., Koprinska, I., Wang, E., Zeng, Y., Li, J. and Kleitman, S.: Detecting Personality Traits Using Eye-Tracking Data, *CHI*, pp. 1–12 (2019).
[6] Brown, B. K. and Campion, M. A.: Biodata phenomenology: Recruiters' perceptions and use of biographical information in resume screening., *J. Applied Psychology*, Vol. 79, No. 6, pp. 897–908 (1994).
[7] Charleer, S., Gutiérrez, F. and Verbert, K.: Supporting job mediator and job seeker through an actionable dashboard, *IUI*, pp. 121–131 (2019).
[8] Cole, M. S., Rubin, R. S., Feild, H. S. and Giles, W. F.: Recruiters' Perceptions and Use of Applicant Résumé Information: Screening the Recent Graduate, *Applied Psychology*, Vol. 56, No. 2, pp. 319–343 (2007).
[9] Côté, S. and Moskowitz, D. S.: On the dynamic covariation between interpersonal behavior and affect: Prediction from neuroticism, extraversion, and agreeableness., *J. Personality and Social Psychology*, Vol. 75, No. 4, pp. 1032–1046 (1998).
[10] DeYoung, C. G.: Cybernetic Big Five Theory, *J. Research in Personality*, Vol. 56, pp. 33–58 (2015).
[11] DeYoung, C. G., Quilty, L. C. and Peterson, J. B.: Between facets and domains: 10 aspects of the Big Five., *J. Personality and Social Psychology*, Vol. 93, No. 5, pp. 880–896 (2007).
[12] Esterwood, C., Essenmacher, K., Yang, H., Zeng, F. and Robert, L. P.: A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction, *CHI*, pp. 1–18 (2021).
[13] Gilbert, P. and Allan, S.: Assertiveness, submissive behaviour and social comparison, *British Journal of Clinical Psychology*, Vol. 33, No. 3, pp. 295–306 (1994).
[14] Golbeck, J., Robles, C. and Turner, K.: Predicting personality with social media, *CHI EA*, pp. 253–262 (online), DOI: 10.1145/1979742.1979614 (2011).
[15] Gosling, S. D., Rentfrow, P. J. and Swann, W. B.: A very brief measure of the Big-Five personality domains, *J. Research in Personality*, Vol. 37, No. 6, pp. 504–528 (2003).
[16] Hofmann, H.: Statlog (German Credit Data), UCI Machine Learning Repository (1994).
[17] Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B. and Ghosh, J.: Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems, pp. 1–19 (2019).
[18] Kanamori, K., Takagi, T., Kobayashi, K. and Arimura, H.: DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization Kentaro, *IJCAI*, pp. 2855–28 (2020).
[19] Kanamori, K., Takagi, T., Kobayashi, K., Ike, Y., Uemura, K. and Arimura, H.: Ordered Counterfactual Explanation by Mixed-Integer Linear Optimization, *AAAI*, Vol. 35, No. 13, pp. 11564–11574 (2021).
[20] Karimi, A.-H., Barthe, G., Balle, B. and Valera, I.: Model-Agnostic Counterfactual Explanations for Consequential Decisions, *AISTATS*, pp. 895–905 (2020).
[21] Karimi, A. H., Barthe, G., Schölkopf, B. and Valera, I.: A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations, *ACM Computing Surveys*, Vol. 55, No. 5 (2022).
[22] Karimi, A.-H., Schölkopf, B. and Valera, I.: Algorithmic Recourse: from Counterfactual Explanations to Inter-

ventions, *FAccT*, pp. 353–362 (2021).
[23] Karimi, A. H., von Kügelgen, J., Schölkopf, B. and Valera, I.: Algorithmic recourse under imperfect causal knowledge: A probabilistic approach, *NeurIPS* (2020).
[24] Keane, M. T., Kenny, E. M., Delaney, E. and Smyth, B.: If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques, *IJCAI*, pp. 4466–4474 (2021).
[25] Keane, M. T. and Smyth, B.: Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI), *ICCBR*, pp. 163–178 (2020).
[26] Kirfel, L. and Liefgreen, A.: What If (and How ...)? - Actionability Shapes People's Perceptions of Counterfactual Explanations in Automated Decision-Making, *ICML-21 Workshop on Algorithmic Recourse* (2021).
[27] Kuhl, U., Artelt, A. and Hammer, B.: Keep Your Friends Close and Your Counterfactuals Closer: Improved Learning From Closest Rather Than Plausible Counterfactual Explanations in an Abstract Setting, *FAccT*, pp. 2125–2137 (2022).
[28] Leavitt, M. L. and Morcos, A.: Towards falsifiable interpretability research (2020).
[29] Leising, D., Sporberg, D. and Rehbein, D.: Characteristic Interpersonal Behavior in Dependent and Avoidant Personality Disorder can be Observed Within Very Short Interaction Sequences, *J. Personality Disorders*, Vol. 20, No. 4, pp. 319–330 (2006).
[30] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence*, Vol. 267, pp. 1–38 (2019).
[31] Mothilal, R. K., Sharma, A. and Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations, *FAT**, pp. 607–617 (2020).
[32] Naumann, P. and Ntoutsi, E.: Consequence-Aware Sequential Counterfactual Generation, *ECMLPKDD*, pp. 682–698 (2021).
[33] Orji, R., Nacke, L. E. and Di Marco, C.: Towards Personality-driven Persuasive Health Games and Gamified Systems, *CHI*, pp. 1015–1027 (2017).
[34] Pawelczyk, M., Broelemann, K. and Kasneci, G.: On Counterfactual Explanations under Predictive Multiplicity, *UAI*, Vol. 124, pp. 809–818 (2020).
[35] Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T. and Flach, P.: FACE: Feasible and Actionable Counterfactual Explanations, *AIES*, pp. 344–350 (2020).
[36] Read, S. J. and Marcus-Newhall, A.: Explanatory coherence in social explanations: A parallel distributed processing account., *J. Personality and Social Psychology*, Vol. 65, No. 3, pp. 429–447 (1993).
[37] Russell, C.: Efficient Search for Diverse Coherent Explanations, *FAT**, pp. 20–28 (2019).
[38] Scheier, M. F. and Carver, C. S.: Optimism, coping, and health: Assessment and implications of generalized outcome expectancies., *Health Psychology*, Vol. 4, No. 3, pp. 219–247 (1985).
[39] Schulenberg, K., Li, L., Freeman, G., Zamanifard, S. and McNeese, N. J.: Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality, *CHI*, pp. 1–17 (2023).
[40] Sefidgar, Y. S., Seo, W., Kuehn, K. S., Althoff, T., Browning, A., Riskin, E., Nurius, P. S., Dey, A. K. and Mankoff, J.: Passively-sensed Behavioral Correlates of Discrimination Events in College Students, *ACM Hum.-Comput. Interact.*, Vol. 3, No. CSCW, pp. 1–29 (2019).
[41] Song, D., Wang, Z., Huang, Y., Ma, L. and Zhang, T.: DeepLens: Interactive Out-of-distribution Data Detection in NLP Models, *CHI*, pp. 1–17 (2023).
[42] Thagard, P.: Explanatory coherence, *Behavioral and Brain Sciences*, Vol. 12, No. 3, pp. 435–467 (1989).
[43] Thompson, E. R.: Development and Validation of an International English Big-Five Mini-Markers, *Personality and Individual Differences*, Vol. 45, No. 6, pp. 542–548 (2008).
[44] To, A., Sweeney, W., Hammer, J. and Kaufman, G.: "They Just Don't Get It": Towards Social Technologies for Coping with Interpersonal Racism, *ACM on Hum.-Comput. Interact.*, Vol. 4, No. CSCW1, pp. 1–29 (2020).
[45] Umstattd, M. R., McAuley, E., Motl, R. W. and Rosengren, K. S.: Pessimism and Physical Functioning in Older Women: Influence of Self-Efficacy, *J. Behavioral Medicine*, Vol. 30, No. 2, pp. 107–114 (2007).
[46] Ustun, B., Spangher, A. and Liu, Y.: Actionable Recourse in Linear Classification, *FAT**, pp. 10–19 (2019).
[47] Van Looveren, A. and Klaise, J.: Interpretable Counterfactual Explanations Guided by Prototypes, *ECMLP-KDD*, pp. 650–665 (2021).
[48] Venkatasubramanian, S. and Alfano, M.: The philosophical basis of algorithmic recourse, *FAT**, pp. 284–293 (2020).
[49] Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P. and Shah, C.: Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review (2020).
[50] Völkel, S. T., Schödel, R., Buschek, D., Stachl, C., Winterhalter, V., Bühner, M. and Hussmann, H.: Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach, *CHI*, pp. 1–14 (2020).
[51] Wachter, S., Mittelstadt, B. and Russel, C.: Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, *Harvard Journal of Law & Technology*, Vol. 20, No. 3, pp. 842–887 (2018).
[52] Wang, Y., Venkatesh, P. and Lim, B. Y.: Interpretable Directed Diversity: Leveraging Model Explanations for Iterative Crowd Ideation, *CHI*, pp. 1–28 (2022).
[53] Wang, Z. J., Vaughan, J. W., Caruana, R. and Chau, D. H.: GAM Coach: Towards Interactive and User-centered Algorithmic Recourse, *CHI* (2023).
[54] Williams, J. B. W. and Benjamin, L. S.: *The Structured Clinical Interview for DSM-IV axis II Personality Disorders (SCID-II)*, American Psychiatric Publishing, Inc., Washington D.C. and London, England, first japanese edition 2002 by igaku-shoin ltd., tokyo edition (1997).
[55] Xie, J., Lipford, H. and Chu, B.-T.: Evaluating interactive support for secure programming, *CHI*, pp. 2707–2716 (2012).
[56] Yeh, I.-C.: Default of Credit Card Clients, UCI Machine Learning Repository (2016).
[57] 小塩真司，阿部晋吾，Cutrone, P.：日本語版 Ten Item Personality Inventory（TIPI-J）作成の試み，パーソナリティ研究， Vol. 21, No. 1, pp. 40–52 (2012).
[58] 外山美樹：楽観・悲観性尺度の作成ならびに信頼性・妥当性の検討，心理学研究，Vol. 84, No. 3, pp. 256–266 (2013).