AIの判定に対する初期の受容態度がその後のアルゴリズミック・リコースの認識に及ぼす影響

冨永 登夢[1,a)]　山下 直美[2]　倉島 健[1]

概要：アルゴリズミック・リコースは AI から不利な判定を下された個人に対して判定理由の理解と将来行動の促進を目的に反実仮想的な提案を示す．既存研究の多くは合理的で実行可能なリコースの生成に注力してきたが，AI から判定を下された時点で (リコースが提供される前に) 被意思決定者がその判定結果を受け入れているか否かは考慮されてこなかった．そこで我々は，判定結果に対する最初の受容態度がその後に提供されるリコースに対する認識を左右し，最終的な受容態度に影響を与えるのではないか，という仮説を検証すべく，自動車ローン審査を題材とするシナリオベースのユーザ実験 (N=534) を行った．統計解析の結果，最初の受容態度が否定的な被験者はリコースの合理性と実行可能性を低いと感じ，最終的に AI の判定結果に対して否定的な態度を示すことが分かった．一方で，リコースが審査基準を説明するもしくは現実的な範囲で実行できる計画を示すと認識されると，被験者の受容態度は否定的から肯定的に変化することが定性分析から確認された．これらを踏まえて，望まない判定結果を下された個人に納得感をもたらすリコース生成システムに関する将来研究の目指すべき方向性や課題について論じた．

1. はじめに

説明性の高い AI 技術の需要が高まる中，アルゴリズミック・リコースが有力な手法として注目されている．アルゴリズミック・リコースは，望まない判定を AI から下された個人に反実仮想的な行動計画を提供することで，判定が下された理由と今後望ましい判定結果を得るために必要な行動を対象者に示す [36]．例えば，ローンの審査に落ちた人に対して “もしあなたの年収が 100 万円増えればローン審査に承認されます” といった説明を行う．その究極的な目的は望ましくない不利な判定結果を受け取った個人を救済することにある [15]．

既存研究は対象者にとって判定結果を納得して受け入れることのできるリコースを提供することに焦点を当ててきた．例えば，ユーザにとって好ましいリコースを調査した研究は，ユーザは直感的・論理的でないリコースに対してさらなる説明を求め [34]，自分の裁量で実行できる範囲の変更を提案するリコースを好む [38] と述べた．工学的な領域では，内容が矛盾しない複数のリコースを生成する手法 [24] や，リコースにおいて提案される行動をこなす順序を導出する手法 [14] など，数多くの技術的解法が提案されている．これらの根底には，説明として合理的であり行動計画として実行可能なリコースを作れば本人は AI の判定結果を受容する，という広く支持される前提がある [15]．

しかし，被意思決定者が (リコースを提示されるより前に) そもそも AI に下された判定結果に納得しているかどうかによって最終的な受容態度は変わる可能性がある．AI による自動的意思決定に関する研究 [1], [27], [28], [29] によれば，AI に対して一度否定的な印象を持つとその後の AI に対する信用や依存度は下がる [23], [31]．これは，たとえ AI の性能が十分高精度だとしても同様である [7], [8]．これらを踏まえ我々は，AI の判定結果に対する最初の受容態度が肯定的か否定的かによってその後のリコースに関する認識が左右され，最終的な受容態度にまで波及効果を及ぼすのではないかと予想した．そこで本研究は合理性と実行可能性をリコースの認識の側面として着目し，まず以下の研究課題を検証した．**RQ1: AI の判定に対する初期の受容態度はその後に AI から提供されるリコースの合理性と実行可能性に関する認識にどのような影響を与えるか？**

次に，リコースに対する認識が AI の判定結果に対する最終的な受容態度に与える影響を調べた．既存研究では，被意思決定者にとって合理的で実行可能なリコースは，対象者が判定結果に納得するような説明を提供できると暗に仮定されている [15]．しかし，合理性や実行可能性に関する

[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.
(現在，京都大学大学院情報学研究科)
[a)] tomu.tominaga@ntt.com

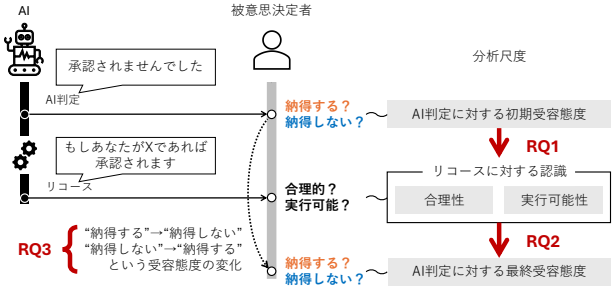


図 1 本研究における研究課題の焦点

認識が実際に受容態度にどのような影響を及ぼすかは実証されていない．そこで以下の研究課題を設定し，リコースに対する認識と最終的な受容態度の関係を明らかにした．**RQ2: リコースに対する合理性と実行可能性に関する認識は AI の判定に対する最終的な受容態度にどのような影響を与えるか？**

最後に，被意思決定者の AI 判定結果への受容態度の変化に寄与したリコースがどのような特徴を持つかを調べた．具体的には，当初は AI 判定に対して否定的だったがその後肯定的になった場合，またその逆の状況が発生した場合，どのようなリコースが提示されたかを調べた．**RQ3: どのようなリコースによって AI の判定に対する受容態度は肯定的から否定的もしくは否定的から肯定的に変化するか？**この結果は，被意思決定者が AI の判定結果に納得できるようにどのようなリコースを作るべきか，もしくは作るべきではないか，を明らかにするのに役立つ．本研究の 3 つの研究課題は図 1 に示す関係にある．

これらを検証するため，我々は自動車ローンを題材としたシナリオベースの被験者実験を行った (N=534)．我々は実験シナリオに現実味を持たせるため，実際に自動車を購入する意欲を持つ被験者が自身のデータを提出し AI に審査されるという手順で実験を進めた．被験者は AI の判定結果に対する受容態度をリコースの提供前後で 2 回行った．また，被験者は提供されたリコースの合理性と実行可能性を評価し，各評価理由を自由記述式で回答した．

統計的解析により，我々は初期受容態度がその後のリコースに対する認識や判定結果に対する最終受容態度に影響を与えていることを確認した．具体的には，初期態度が否定的であるほど，リコースの合理性と実行可能性を低く評価する傾向にあり，最終態度も否定的になることが分かった．また，定性的分析により受容態度を変化させるリコースの特性を明らかにした．例えば，否定的から肯定的に変化するとき，リコースは判定結果の審査基準や対象者の欠点を明瞭に説明している，難しすぎず簡単すぎない行動計画を提案している，と認識されていることが分かった．反対に，リコースが公平性や実世界文脈に配慮していないと認識されると，受容態度が肯定的から否定的に変わることも確認した．我々の結果はリコースによって AI から望まぬ判定を受けた個人に説明や推薦を提供するプロセスにおける初期受容態度の役割を解明し，AI 判定に対する受容を促進するための効果的なリコース生成について重要な実証的知見を提供するものである．

本研究の貢献は以下の通りである．

(1) リコース提供前に AI の判定結果に納得していない被験者は，その後に提供されるリコースの合理性と実行可能性を低く認識する傾向にあり，結果的に最終的な受容態度が否定的になることを突き止めた．
(2) 受容態度を変化させるリコースの持つ特性を見出した．リコースが判定結果を正当化する，審査基準を説明する，現実的な範囲で適度な難易度の計画を示す，個人の生活や公平性に配慮することが重要な要因であった．
(3) AI から不利な判定を下された被意思決定者が最終的にその判定を受容できるようなリコースを生成するシステムの設計指針と将来課題を示した．

2. 関連研究

信用審査 [19]，医療診断 [3]，法的裁定 [4] など，リスクの高い領域において特定の基準にしたがって AI や機械学習モデルが評価や査定を下す自動的な意思決定を人間がどう認識しているかが近年注目されている．特に，AI-assisted decision making や Human-AI collaboration の分野で，人間と信頼関係を築くプロセス [13], [22], [40] やアルゴリズムの性能の低さの影響 [23] が調査されている．例えば，エラーやバイアスなどに起因して AI システムの挙動に対する第一印象が悪い場合，ユーザはその後システムに対する信用レベルを著しく低下させ [31]，AI の意思決定支援への依存を避ける傾向にある [7], [23]．一方で，AI が高精度に機能しているという印象を最初に持つと，システムからの提案をしばしば過剰なほどに受け入れるようになることが分かっている [25], [26]．

これらの研究は意思決定者 (意思決定を下す人) を対象としたのに対し，我々はアルゴリズミック・リコースの文脈における被意思決定者 (意思決定を下される人) に注目する．我々は上述の既存研究の結果から，最初に被意思決定者が AI の判定結果に対して否定的な態度を持つ場合，その後に提供されるリコースへの評価や最終的な判定結果に対する受容態度が悪化するのではないかと予想した．アルゴリズミック・リコースの本来の目的は AI から不利な判定を下された個人の救済にあるため，この検証によって，そもそも判定結果を不服に思う被意思決定者の態度がその後のプロセスにおいて果たす役割を追求することは重要である．

既存のアルゴリズミック・リコースに関する研究は，判定理由の説明として合理的であり [9], [12], [36]，行動計画として実行可能である [16], [17] リコースを生成することに専念してきた．その多くが，リコースの合理性や実行可

能性を測定する目的関数を定め，最適化する手法を提案している [15], [35]．このような技術的な研究だけでなく，リコースに対する被意思決定者の実際の反応を観測する被験者実験も行われている．例えば，AI システムに対する被意思決定者の理解を促進するリコースの提示方法を調べた研究によれば，被意思決定者は自動的意思決定システムが正確かつ一貫して人間の推論を忠実に再現することを期待する [10] ため，直感に反するもしくは論理的でないリコースが提示されるとさらなる説明を要求したことを確認している [34]．これに対し，Wang ら [38] は，被意思決定者が自身の好みに合うリコースをインタラクティブに選択することのできるインタフェースを開発し，ローン審査に関するリコースに対するユーザの嗜好を利用ログとアンケートの回答データから探索的に調べた結果，自分の裁量で実行しやすいリコースが選ばれる傾向にあると報告した．これらの結果は，被意思決定者にとって合理的で実行可能なリコースが好まれやすいことを示唆している．

しかし，これらの研究の根本にある，合理的で実行可能なリコースによって被意思決定者は AI の判定結果を受容するという前提はこれまで検証されていない．本研究はこの前提に対する実証的な知見を提供する．

3. 実験

本研究は，被意思決定者の AI の判定結果に対する受容態度とリコースに対する認識の関係を調査するため，自動車ローン申請のシナリオを題材としたオンライン調査を行った．先行研究 [32] と同様に，具体的には以下のシナリオを設定した．

> 被験者は，自身の年収の 3 分の 1 に相当する自動車ローンを融資してもらいたい．そこで彼らは金融機関を訪れ，プロファイルデータを提出して審査を受けることにした．そこでは AI システムが内部基準にしたがってローン申請を審査している．審査の結果，被験者のローン申請は不承認という結果となった．被験者は，不承認となった判定結果の理由を理解し，今後承認を得るために必要な行動を把握するため，AI が生成した複数のリコースを確認することにした．

ここでは，上記金融機関は年収の 1/4 を超える融資申請を全て却下するという内部規則にしたがって審査していると仮定している．そのため，全ての被験者の申請は却下されるが，この方針は実験中に被験者には開示されない．

上記のシナリオをもとに，我々は 2024 年 2 月 7 日から 3 月 21 日の期間で本実験を実施した．手順は図 2 に示される通りである．それぞれの詳細を以降の節に記す．なお，本実験はパブリックヘルスリサーチセンターの倫理審

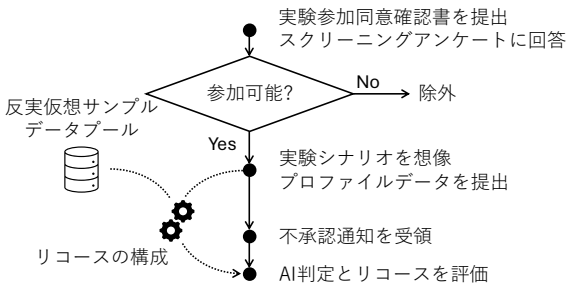


図 2 全体的な実験手順

査委員会[*1]により審査および承認されている (承認番号：PHRF-IRB 24A0002)．全ての実験手続きはここで提供されたガイドラインにしたがって実施された．

3.1 被験者の募集と選定

被験者はオンラインリサーチ会社 ASMARQ[*2]を通じて募集された．本実験に参加可能な被験者の条件は，(1) 現在，民間企業または公的機関に勤務していること，(2) 車の購入に興味があること，(3) 参加時にローンを保有していないこと，(4) 年収が 1000 万円未満であること，の 4 つである．これらは本シナリオを採用した先行研究と同様 [32] である．我々は応募者の実験参加同意確認の意思を確認した後，スクリーニングアンケートによって彼らがこれらの条件を満たすかどうかを確認した．

スクリーニングアンケートを行った結果，最終的に 550 名が被験者として選定された．男性は 402 名 (73.1%)，女性は 148 名 (26.9%)，全体の平均年齢は 48.9±10.3 歳であった．本実験における全ての調査を完了した被験者には 600 円相当の報酬が支払われた．

3.2 プロファイルデータの提出

スクリーニングアンケート終了後，実験参加資格を持つ被験者はシナリオ実験に進む．ここではまず被験者は前述のシナリオを想像するように指示される．その後，シナリオ上の審査で必要なプロファイルデータの提出を依頼される．プロファイルデータの項目と選択肢を表 1 に示す．これらの項目は先行研究 [32] と同様である．

3.3 リコースの構成

我々は被験者から提出されたプロファイルデータに基づいて各被験者に 5 つのリコースを構成した．最終的に被験者に提示されるリコースの例を表 2 に示す．

表 2 に示されるように，リコースは一般に被意思決定者のプロファイルデータとその対比となる別のプロファイルデータとの差分を被意思決定者に教示する．ここで被験者のプロファイルデータを入力サンプル x，対比と

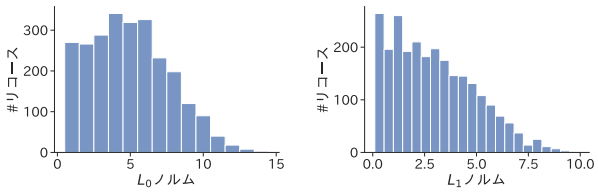
[*1] https://www.phrf.jp/rinri/
[*2] https://www.asmarq.co.jp/global/

表 1 審査用プロファイルデータ項目 (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
1	居住地	1. 東京 / 2. 東京以外	
2	居住形態	1. 持ち家 / 2. 賃貸・寮	
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)	$x_3^* \geq x_3$
4	勤務先	1. 一般企業 / 2. 公的機関	
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	$x_5^* \geq x_5$
6	勤続年数 (年)	1. 0-1 / 2. 1-3 / 3. 3-5 / 4. 5-10 / 5. 10-20 / 6. 20-	$x_6^* \geq x_6$
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-	$x_7^* \geq x_7$
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-	
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-	
10	副業数	1. なし / 2. 1 / 3. 2 / 4. 3 / 5. 4 / 6. 5-	
11	転職歴	1. なし / 2. あり	$x_{11}^* \geq x_{11}$
12	海外勤務歴	1. なし / 2. あり	$x_{12}^* \geq x_{12}$
13	海外留学歴	1. なし / 2. あり	$x_{13}^* \geq x_{13}$
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	$x_{14}^* \geq x_{14}$
15	Facebook 利用	1. 未登録・友達なし / 2. 登録済・友達あり	
16	LinkedIn 利用	1. 未登録・友達なし / 2. 登録済・友達あり	

表 2 被意思決定者に提供されるリコースの例 (実験時 “もしあなたが右のようなプロファイルの保有者であれば審査に通過します” と被験者に説明)

	☹ 現状		☺ 理想
...	...		...
勤務先	一般企業		一般企業
職位	一般社員	→	主任
...	...		...



(a) L_0 ノルム　(b) L_0 ノルム

図 3 評価用に構成されたリコースの距離尺度の分布

なる別のプロファイルデータを反実仮想サンプル x^* とする．反実仮想サンプル $x^* \in X$ は，N 次元の特徴量空間 $X = X_1 \times \cdots \times X_N$，特定の条件で定義される X の部分集合 $P(X) \subseteq X$，事前に学習された二値分類 AI モデル (例：ローン審査) $F : X \to \{\text{Approval}, \text{Rejection}\}$，不利な決定を下された入力サンプル $x \in X (F(x) = \text{Rejection})$ が与えられた時，以下の最適化問題を解くことで特定される．

$$x^* = \underset{x, x^* \in P(X)}{\operatorname{argmin}} \; d(x, x^*) \text{ subject to } F(x^*) \neq F(x) \quad (1)$$

ここで目的関数 $d : X \times X \to \mathbb{R}_{>0}$ は，入力サンプルと反実仮想サンプルの差分を定量化する距離尺度を表す．

本実験では，距離尺度 d として L_0 ノルムと L_1 ノルムを用いる．$\mathbb{I}$ を指示関数，M_i を特徴量 i の最大変更範囲として，それぞれ以下のように計算される．

$$L_0 = \sum_i \mathbb{I}[x_i \neq x_i^*], \; L_1 = \sum_i |x_i - x_i^*| / M_i \quad (2)$$

L_0 ノルムは x から x^* に変更される特徴量の数に相当し，L_1 ノルムは x から x^* のマンハッタン距離を示す．

本実験において，入力サンプル x に対応する反実仮想サンプル x^* は既存データセット [32] の中から選択される．ここで反実仮想サンプルは 2 つの要件を満たす必要がある．第一に，x の年収に対して x^* の年収は 4/3 倍以上とする．これはシナリオ実験で仮定されている融資申請額が年収の 1/3 であり，金融機関の内部規則において年収の 1/4 以上の融資申請額は不承認となるため，入力サンプルが承認されるためには年収が 4/3 倍以上となる必要があるためである．第二に，x と x^* は表 1 に示される制約条件を満たすとする．この制約条件により，リコースで提案される変更を現実的に存在しうる範囲におさめる．

これらの条件下で x に対する x^* は以下の手順にしたがって 5 つ選択される：(1) L_0 ノルムが最小のサンプル 1 つ，(2) L_1 ノルムが最小のサンプル 1 つ，(3) 無作為に 3 つ．この選択方法によってさまざまな距離のリコースを構成できる [32]．これらの手続きで構成されたリコースの距離の分布を図 3 に示す．

3.4 AI 判定とリコースの評価

被験者はプロファイルデータの提出後，まず最初にローン申請の不承認通知を受け取る．そこで我々は，被験者に当判定結果に対する受容態度に関する以下の質問に回答してもらい，回答値を初期受容態度として計測した．

初期受容態度　あなたは今回の判定結果に納得できます

か？ (7 段階評価：“全く納得できない”～“非常に納得できる”)

次に我々は，被験者のプロファイルデータに基づいて構成したリコースを 5 つ提示し，各リコースに対して以下の設問への回答を依頼した．これらの回答値をリコース合理性，リコース実行可能性，最終受容態度として計測した．なお，リコース合理性とリコース実行可能性については，評価理由を自由記述式で回答するように被験者に指示した．また，設問文においては，専門用語であるリコースを提案プランと言い換えた．

リコース合理性　あなたは AI システムの提案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妥当だと思いますか？ (7 段階評価：“全く思わない”～“強くそう思う”)

リコース実行可能性　あなたは AI システムの提案プランを実行することをどの程度難しい/易しいと感じますか？ (7 段階評価：“非常に難しい”～“非常に易しい”)

最終受容態度　AI システムの提案プランを見た今，あなたは今回の判定結果に納得できますか？ (7 段階評価：“全く納得できない”～“非常に納得できる”)

リコース実行可能性を回答させる際に，提示されている各項目の変更内容が不可変かどうかも合わせて尋ねた．本実験では，年齢や性別など自身の裁量や努力によって変えることのできない属性をリコースのプロフィール項目に含めていないが，採用された 16 項目 (表 1) が実際に不可変かどうかは被験者に依存する．そこで，被験者に各変更項目の不可変性について直接尋ね，不可変と評価された項目を 1 つでも含むリコースは本分析から除外した．

4. 分析

550 名の被験者から判定結果に対する受容態度とリコース合理性と実行可能性に関する回答データが合計で 2750 件得られた．このうち，リコースの変更内容が不可変だと回答された 231 件のデータを除外し，残りの 2519 件 (534 名分) を用いて分析を進めた．その詳細を以下に述べる．

4.1 初期受容態度がリコース合理性および実行可能性に及ぼす影響

AI 判定に対する初期の受容態度がリコース合理性および実行可能性に与える影響を調査するため，我々は一般化加法モデル (GAM) を用いた．このモデルは，非線形関数を基底関数とするノンパラメトリックな回帰モデルを使用し，説明変数と目的変数の複雑な関係を平滑化スプラインを利用して柔軟に捉えることができる．

RQ1 の分析では，説明変数は初期受容態度，目的変数はリコース合理性もしくは実行可能性である．そのため，ここでは 2 つの異なる分析モデルからそれぞれ平滑化スプライン関数が導出される．それぞれのモデルにおいて F 検定を実施し，初期受容態度がリコース合理性または実行可能性に有意な影響を与えるかどうかを評価した．

4.2 リコース合理性および実行可能性が最終受容態度に及ぼす影響

本分析の目的は，リコース合理性と実行可能性が AI 判定に対する最終的な受容態度に与える影響を明らかにすることである．この分析に必要となるリコース合理性，リコース実行可能性，および最終受容態度の 3 つのデータは，1 つのリコースに対応する形でひとりの被験者から繰り返し測定される．我々は被験者内比較に基づく洞察 (例：被験者はリコースをより合理的または実行可能と認識すると AI 判定をより受け入れやすくなる) を得るため，一般化加法混合モデル (GAMM) を使用した．GAMM は GAM に混合効果パラメータを導入することで被験者個々人に由来する影響を分離できるように GAM を拡張するため，反復測定データから被験者内比較としての結果を導出できる．

この GAMM において目的変数は最終受容態度であり，説明変数はリコース合理性とリコース実行可能性である．RQ1 と同様に F 検定を使用して説明変数が目的変数に与える影響の統計的有意性を確認した．

4.3 受容態度の変化に影響を与えるリコースの特性

RQ3 は，被意思決定者の受容態度が変化 (反転) するようなリコースの特性を特定することを目的としている．ここでは，受容態度が否定的から肯定的または肯定的から否定的に変化した時のリコースに対して被験者がどのような評価理由を報告しているかを探索的に調べた．

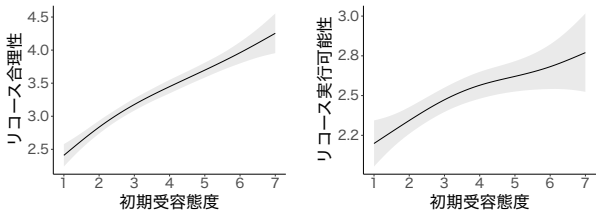
初期および最終受容態度は 7 段階で評価されているため，評価値が 3 以下であれば否定的，5 以上であれば肯定的とした．評価値が 4 の場合は中立と見なし，分析から除外した．その後，受容態度が肯定的から否定的，または否定的から肯定的に変化したリコース評価データを抽出した．前者を PN 群 (肯定的から否定的)，後者を NP 群 (否定的から肯定的) と呼ぶ．PN 群には 307 件，NP 群には 321 件のリコース評価データが含まれた．

リコース合理性および実行可能性の評価理由を尋ねる質問 (“そのように評価した理由をお答えください”) に被験者が自由記述式で回答した内容に対して，我々はテーマティック分析 [5] を用いた質的分析を行った．回答内容からリコース特性に関する個別のコードを帰納的に生成し，それらを用いてテーマを初期化し，反復的に洗練させることでテーマを特定した．

5. 結果

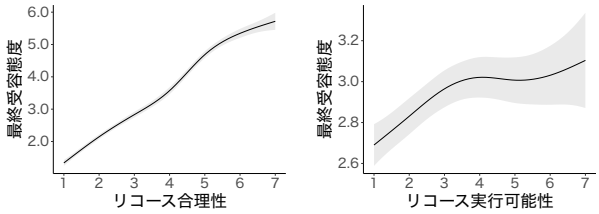
5.1 初期受容態度がリコース合理性および実行可能性に及ぼす影響

一般化加法モデル (GAM) による分析から，初期受容態



(a) 初期受容態度に対するリコース合理性　(b) 初期受容態度に対するリコース実行可能性

図 4 一般化加法モデル (GAM) から得られた，初期受容態度から (a) リコース合理性もしくは (b) リコース実行可能性を説明する平滑化スプライン関数



(a) リコース合理性に対する最終受容態度　(b) リコース実行可能性に対する最終受容態度

図 5 一般化加法混合モデル (GAMM) から得られた，(a) リコース合理性もしくは (b) リコース実行可能性から最終受容態度を説明する平滑化スプライン関数

度がリコース合理性 ($F = 61.08$, $p < 0.001$) および実行可能性 ($F = 9.78$, $p < 0.001$) の両方に統計的に有意な影響を与えることが分かった．図 4 にそれぞれ GAM から得られた平滑化スプライン関数を示す．リコース合理性は，初期受容態度が肯定的であるほど増加し，否定的であるほど減少する．リコース実行可能性も同様に正の相関を示すが，合理性に比べその相関は弱い．初期受容態度のスコア 1 と 7 で平滑化スプライン関数の出力値を比較すると，合理性の関数では 1.85 の差が見られるが，実行可能性の関数では 0.57 の差に留まる．

全体として，初期受容態度はリコース合理性および実行可能性の両方に有意な影響を与え，特に合理性に対する影響がより強い．AI 判定を受け入れにくいと感じる人は，その後のリコースを合理的でないまたは実行しづらいとみなす傾向にあることを示唆している．

5.2 リコース合理性および実行可能性が最終受容態度に及ぼす影響

一般化加法混合モデル (GAMM) より，リコース合理性 ($F = 12.71$, $p < 0.001$) および実行可能性 ($F = 714.62$, $p < 0.001$) が最終受容態度に有意な影響を与えることが確認された．図 5 に GAMM から得られた平滑化スプライン関数をリコース合理性 (図 5(a)) とリコース実行可能性 (図 5(b)) に分けて示す．リコース合理性は最終受容態度と強い相関を示す一方で，リコース実行可能性はその値が低い領域でのみ正の相関を示している．

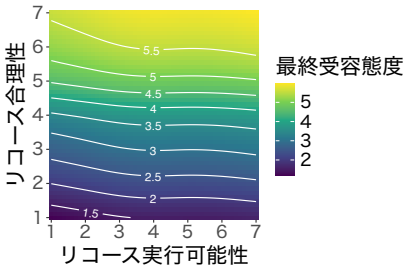


図 6 縦軸をリコース合理性，横軸をリコース実行可能性とする最終受容態度の平滑化スプライン関数 (等高線)

図 6 は，リコース合理性と実行可能性が最終受容態度に同時に与える影響を直観的に理解するため，実行可能性を横軸，合理性を縦軸にとり，GAMM から得られた最終受容態度の平滑化スプライン関数の出力値をプロットした等高線図を示す．この図から，リコース実行可能性よりもリコース合理性の方が最終受容態度を強く支配していることが分かる．例えば，リコース実行可能性が 2 に設定された場合，リコース合理性が高まることで最終受容態度も上昇する．しかし，合理性が 2 に設定された場合，実行可能性が高まっても最終受容態度は大きく上昇しない．

これらの結果は，リコースを合理的とみなすことが AI 判定の最終的な受容に繋がりやすい一方で，リコースを実行しやすいと感じることは必ずしもその後の受容に強く結びつくわけではないことを示している．

5.3 受容態度の変化に影響を与えるリコースの特性

リコース合理性がリコース実行可能性と比べて AI 判定の初期および最終受容態度とより強く相関していることが確認されたため，リコース合理性の評価理由の自由記述回答を分析した．テーマティック分析 [5] を用いた結果，リコース特性は以下の 4 つのテーマに集約された：原因と結果の関係の明確性，判定基準の透明性，実施の実現可能性と望ましさ，および審査と AI に関する懸念．これらのテーマとリコース特性との関係を表 3 に示す．

5.3.1 原因と結果の関係の明瞭性

多くの被験者は，初期受容態度が否定的から肯定的に変わった場合 (NP 群) であってもその逆 (PN 群) であっても，自分が不承認という結果となった原因をリコースから見出そうとしていた．NP 群の被験者は，「勤続年数の多さは重要だと思うから」(P663) や「持ち家もないし、高卒だし課長でもないから」(P696) といった報告にあるように，リコースの変更内容が重要な評価項目 (NP1，84/321) や自分自身の欠点 (NP2，21/321) を説明していると認識し，最終的に肯定的な態度を示した．一方で，リコースで提案される変更が不承認という判断を正当化できていないと感じた場合 (PN1，44/307)，被験者の受容態度は悪化した．彼らは「なぜ」や「理解できない」といった表現で自分の疑問や不満を述べた．例えば「学歴だけで判断は理解

表 3 受容態度を変化させるリコースの特性．ID 列の NP および PN は受容態度が否定的から肯定的に (NP 群) もしくは肯定的から否定的に (PN 群) 変化したグループを指す．

	リコース特性			
テーマ	ID	説明	N	%
原因と結果の関係の明瞭性	NP1	特徴の変更によって判定結果の理由を説明する	84	26.2
原因と結果の関係の明瞭性	NP2	特徴の変更によって対象者の欠点を説明する	21	6.5
審査基準の透明性	NP3	特徴の変更によって審査基準を説明する	74	23.1
審査基準の透明性	NP4	特徴の変更によって対象者が操作できないルールを説明する	15	4.7
計画の実現可能性と望ましさ	NP5	特徴の変更によって現実的な計画を提案する	41	12.8
計画の実現可能性と望ましさ	NP6	特徴の変更によって非現実的な計画を提案する (受容可能)	66	20.6
(その他)			20	6.2
原因と結果の関係の明瞭性	PN1	特徴の変更によって判定結果の理由を説明できない	44	14.3
審査基準の透明性	PN2	特徴の変更によって審査基準を説明できない	80	26.1
審査基準の透明性	PN3	特徴の変更によって判定結果の理由や審査基準を不明瞭にする	9	2.9
計画の実現可能性と望ましさ	PN4	特徴の変更によって非現実的な計画を提案する (受容不可能)	68	22.1
計画の実現可能性と望ましさ	PN5	特徴の変更によって必要性のない計画を提案する	30	9.8
計画の実現可能性と望ましさ	PN6	特徴の変更によって日常生活に合わない計画を提案する	20	6.5
審査と AI に対する懸念	PN7	特徴の変更に対象者が使用してほしくない項目を用いる	36	11.7
審査と AI に対する懸念	PN8	特徴の変更によって AI に対する嫌悪感を引き起こす	6	2.0
(その他)			14	4.6

できない」(P436) や「東京都でないといけない理由が不明」(P345) という記述が見られた．

5.3.2 審査基準の透明性

AI の判定に関連する審査基準がリコースの変更内容から被意思決定者に伝わったかどうかが受容態度に影響を与えていたことが分かった．被験者は，シナリオ内の金融機関によって使用されたであろう審査基準やプロセスをリコースの変更項目から推測でき，かつそれに同意できる時，当初は納得していなくても最終的に AI の判定を受け入れた (NP3，74/321)．本実験ではローン申請の審査基準は被験者に知らされていないが，このような被験者の自己報告には，収入，返済能力，信用といった審査基準に関連しうる用語がしばしば言及されるという特徴が見られた．例えば「勤務時間を長くすることで収入を増やせば返済能力が高まると言う考え方は納得がいくから」(P498) や「信頼度が高くなっているから」(P694) という記述が見られた．また，リコースが自分ではコントロールできない既定のルールに則っていると認識した時にも受容態度は改善された (NP4，15/321)．しかし，この肯定的な態度は「決め事と言われてしまえば反論出来ない」(P742) など，承認される可能性への諦めから来ていた．

一方で，受容態度が悪化した被験者の一部は，リコースから提案される変更を確認した際に「学歴や勤務時間はローンとの関連性がないから」(P087) や「具体的な内容を見ると，安定している事が評価になっていない事もあり，なぜ不承認なのかより分からなくなったので」(P531) など，審査基準が不適切で曖昧であると感じていた (PN2，80/307；PN3，9/307)．

5.3.3 計画の実現可能性と望ましさ

行動計画としてのリコースに関連する特性が確認された．具体的には，被験者がリコースで提案される変更内容を現実的な範囲内で実行できると見なした場合，受容態度は改善された (NP5，41/321)．このような被験者は「努力」や「もう少し」といった表現を用いて行動する意欲を示し，「審査に通らなかったものの、もう少しで通りそうという可能性を感じるから」(P126) や「努力で挽回できる範囲だと感じるから」(P836) などと述べた．一方，提案される変更内容が過度に単純であると認識されるリコースは被験者の受容態度を悪化させた (PN5，20/307)．彼らは「大差ない違いだから」(P077) や「大卒で十分だと思うから」(P015) などを納得できない理由として挙げた．

リコースがあまりにも現実的でない変更を提示した時，被験者の反応は 2 つに分かれた．一方の被験者は諦めから判定結果を受け入れた (NP6，66/321) が，もう一方は抵抗的な態度を示した (PN4，68/307)．前者は，提示されたハードルの高さに自身の無力感を覚え「承認とかけ離れている項目があるから」(P037) と述べた．後者は，リコースの計画が「現実的とは思えない」(P212)「ごく限られた人にしか承認されない提案」(P911) と考え，不服感を示した．

被験者はリコースの提案が自分たちの生活，仕事，社会環境といった個人的な状況において現実的に実行可能かどうかにも注意していた (PN6，20/307)．例えば，P751 は「現状の日常生活は様々なことがつながって成り立っている．安易に一部分だけを変更できない」と述べた．また，法的な制約や職務上の義務がリコースの計画の実施に影響を与えることもあり，「1 日の勤務時間が提示された内容では違法だから」(P542) という報告例も見られた．たとえリ

コースの内容が単純だとしても対象者の現状や環境における制約と衝突してしまう場合，受容態度は悪化する．

5.3.4 審査と AI に対する懸念

リコースの提案する変更が被験者にとって不利または不都合な場合に受容態度は悪化した (PN7，36/307)．特に，リコースが不公平または差別的だと認識すると被験者は AI の判定に対して懸念を表明し，「居住地や転職経験などでの審査は不平等であると感じるから」(P765) や「出身地による差別だと思うから」(P958) などと報告した．彼らは職務 (P468)，学歴 (P468)，居住地 (P958，P765)，転職経験 (P765) などの項目に対してこのような懸念を示した．

また，一部の被験者はリコースの内容ではなく，AI そのものに対して嫌悪感を示した (PN8，6/307)．例えば P082 は「AI に言われても納得出来ないから」と回答した．彼らの初期受容態度は肯定的であることを踏まえると，AI に対する嫌悪感は元々のものではなくリコースを確認した後に生じた可能性が高い．しかし，我々は被験者の自己報告からその理由を明らかにできなかった．この問題については今後の課題とする．

6. 考察

6.1 結果の解釈

否定的な初期受容態度がリコースの評価の悪化につながるという結果 (5.1 節) は先行研究の知見から解釈できる．過去の研究では，AI の不完全な性能に対して否定的な印象を持つとその後 AI からのアドバイスを受け入れにくくなることが報告されている [7], [23], [31]．これは，AI や機械学習モデルが人間のように合理的に考えかつ一貫して正確に機能するという期待を満たせないことがユーザ満足度の低下を引き起こすためである [30], [34], [39]．本実験における被験者の否定的な初期受容態度は，自動車ローン申請に承認されるだろうという自身の期待に反して不承認の判定を下した AI に対する不満やフラストレーションに起因している可能性がある．その結果，彼らは AI が提供したリコースを否定的に捉えたと考えられる．

また，リコースに対する評価の悪化が否定的な最終受容態度と関連するという結果 (5.2 節) は，リコース合理性やリコース実行可能性と受容態度の間に正の相関があるとする既存研究の暗黙的な仮定と一致する [15], [36]．本研究はこれを裏付ける実証的な知見を提供する位置付けにある．さらに重要なこととして，合理性と実行可能性の双方が最終受容態度に統計的には関連している一方で，合理性の影響が圧倒的に大きく実行可能性の影響は限定的であった．この結果は，被意思決定者の肯定的な受容態度を促進するためには，リコースの実行可能性よりも合理性を担保するほうが優先的であることを示唆する．

これらの定量的な解析により，統計的には被意思決定者の初期受容態度はその後のリコース合理性や実行可能性の認識さらには判定結果に対する最終的な受容態度を左右する役割を果たすことが明らかとなった．つまり，初期受容態度が否定的であれば合理性や実行可能性の認識は悪化し，否定的な最終受容態度につながると言える．

このような統計的な傾向が見られる一方で，リコースがある程度の努力を要する実現可能な計画を提示している，現実世界の文脈や使用項目の公平性に配慮している，もしくは審査基準や被意思決定者の不足点を明確に説明していると認識されると，被意思決定者の受容態度は否定的から肯定的に変化することが定性的分析から明らかになった (5.3 節)．被験者は非現実的で挑戦的すぎるもしくは取るに足らない単純すぎるリコースを嫌厭する一方で，現実的なハードルを提供するリコース，すなわち，簡単すぎずかつ困難すぎない計画を薦めるリコースを好んだ (5.3.3 節)．これは，被意思決定者がリコースを行動計画にしたがう意欲に基づいて評価し，その意欲が最終的に判定を受け入れるかどうかに影響を与えることを示唆している．多くの既存研究 [18], [33], [34], [38] において実行可能性に焦点が当てられる中，リコースが対象者の現状を改善するための介入手段となる可能性も指摘されている [15]．リコースによる介入を成功させるためには，実行意欲が受容態度に与える影響を探ることが将来的に重要である．

本実験では，生活，仕事，社会的規範といった現実世界の文脈 (5.3.3 節) やリコースに使用される項目の公平感 (5.3.4 節) など，個人的な視点と密接に紐づくリコース特性が確認された．これは本研究の実験デザインによるものである．従来の研究 [11], [20], [21], [38] では，被験者が実験シナリオ内の架空の登場人物，すなわち自分以外の誰かの役割を担って AI から提供される説明を評価していたが，本研究の被験者は実際に車の購入に関心があり，彼ら自身のプロファイルデータを提出し，不承認通知を受け取った後，リコースを確認している．このように現実的状況をより忠実に再現することで，公平性の認識や日常生活との親和性などの個人的な要因と受容態度の関連を明らかにした．本研究は個人の文脈を考慮したリコースの必要性 [2], [35] を実証的に確認したものであり，どのような個人の文脈が関連するかを具体的に示した．

判定結果を正当化し自身の欠点を教示するような妥当性の高いリコースや，判定基準を説明するような透明性の高いリコースが受容態度の向上に寄与するという結果 (5.3.1 および 5.3.2 節) は既存研究と一貫する．この結果は，AI モデルの出力結果が人間の期待から逸脱しておらず [34] かつその内部プロセスが開示される場合 [28]，その AI モデルは対象者に好まれるという先行研究によって支持される．

6.2 将来研究の方向性と課題

本研究の結果から，被意思決定者が判定結果に納得できるようなリコースの要件は，社会生活や公平感に配慮し

た上で，判定結果を正当化し，審査基準を透明化し，現実的な範囲で適度な難易度の行動計画を推奨することであると言える．このような極めて主観的な性質を備えるリコースを生成するためには，距離関数の最適化によって反実仮想サンプルを特定する既存の “stand-alone” なアプローチ [17], [35] ではなく，被意思決定者との対話を通じて彼らの嗜好を検出しリコース生成プロセスに組み込む “interactive” なアプローチが必要である．例えば Wang らが設計したインタフェース [38] のように，被意思決定者がリコースの使用項目やその値を指定することで様々なオプションを探索できるシステムであれば，自身の嗜好に合うリコースを柔軟に設定できる．

しかし，判定結果に納得するという目的においてどの程度のインタラクティブ性が最適かは今後検証される必要がある．一般に，もし原因が X ではなく X’ だったら結果は Z ではなく Z’ であっただろう，という反事実的思考は認知資源を多く消費するだけでなく，後悔，罪悪感，自己嫌悪といった負の感情と結びつきやすい [6]．そのためインタラクティブ性を高めすぎると被意思決定者が数多くのリコースを目にすることになり，認知的にも心理的にも負担を強いる可能性がある．実際，本実験の被験者 P431 は，提示されたリコースについて「プランというと計画ですが、これは過去のことをできてないと非難されていると感じます」と報告した．この点を考慮すると，判定結果に納得できるリコースを生成する計算システムを将来的に確立するためには，被意思決定者の認知的および心理的負担を軽減しつつ彼らとの対話を通じて嗜好を効果的に検知することが重要な研究課題となる．

また，我々はリコースに対する認識としてリコースの合理性と実行可能性に注目したが，本実験結果からリコースの提案内容を実行しようとする意欲も受容態度に関連する重要な要素であることが分かった．将来的には被意思決定者の実行意欲を引き出す仕組みを考案することが重要であると言える．そのための第一歩として，被意思決定者がどのようなリコースによって意欲的になるかを将来的には調査する必要がある．

6.3 本研究の制約事項

本研究におけるシナリオ実験は先行研究 [32] を参考に自動車ローンを題材とした．自動車の購入は多くの人にとって馴染みがあり，高額で，現実に即したテーマである [32]．このような性質から大きく乖離したテーマにおいては本研究と異なる結果となることが予想されるため，同様の結果が確認されるかどうかは今後検証が必要である．

また，被験者に女性が少なく (26.9%)，性別の分布に偏りがある．2023 年の東京都政策企画局による調査 [41] では，週に 1〜2 回以上の運転をする女性は対象者のうち 31.3% であった．これを考慮すると，本実験は自動車の購入に関心のある人を対象としたことで，女性の割合が低くなったと考えられる．既存研究によれば，自身にとって不利な判定が AI から下された時，女性の方が男性よりも不公平性に気付きやすい [37]．そのため，女性被験者の少ない本実験において不公平感に関する自己報告数は過小評価されている可能性がある．この点については今後検証予定である．

7. 結論

我々は 550 名を対象とした被験者実験により，AI の判定結果に対する初期受容態度がその後のリコースの認識や最終受容態度を支配する極めて重要な役割を果たすことを実証した．初期受容態度が否定的であるとリコースに対する認識が悪化し，それが否定的な最終受容態度を招く傾向にあった．これは，AI から望ましくない判定を下された個人の救済を目的とするアルゴリズミック・リコースの実現が挑戦的な問題であることを意味する．さらに我々は，リコースが判定結果を正当化し，審査基準を説明し，公平性や個人の文脈に配慮し，現実的な範囲で実現可能な計画を提案した場合，被意思決定者の受容態度は否定的から肯定的に反転することを確認した．これは上記の問題に対する一つの解を示している点で重要である．これらを踏まえ，反事実的思考に由来する認知的・心理的負担を抑えながら被意思決定者の嗜好をインタラクティブに捉えることで個別化されたリコースを生成し受容態度を改善するシステムの設計指針と将来課題を論じた．我々は，本研究成果が否定的な影響を受けた個人をリコースによって救うためのインタラクションシステムの基盤となり，人間と AI による意思決定の分野の発展に貢献することを切に願う．

参考文献

[1] Ahn, D., Almaatouq, A., Gulabani, M. and Hosanagar, K.: Impact of Model Interpretability and Outcome Feedback on Trust in AI, *Proc. of CHI*, pp. 1–25 (2024).
[2] Barocas, S., Selbst, A. D. and Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons, *Proc. of FAccT*, pp. 80–89 (2020).
[3] Begoli, E., Bhattacharya, T. and Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making, *Nature Machine Intelligence*, Vol. 1, No. 1, pp. 20–23 (2019).
[4] Berk, R. A., Sorenson, S. B. and Barnes, G.: Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions, *J. of Empirical Legal Studies*, Vol. 13, No. 1, pp. 94–115 (2016).
[5] Braun, V. and Clarke, V.: Using thematic analysis in psychology, *Qualitative Research in Psychology*, Vol. 3, No. 2, pp. 77–101 (2006).
[6] Byrne, R. M.: Counterfactual Thought, *Annual Review of Psychology*, Vol. 67, No. 1, pp. 135–157 (2016).
[7] Dietvorst, B. J., Simmons, J. P. and Massey, C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err, *J. of Experimental Psychology: General*, Vol. 144, No. 1, pp. 114–126 (2015).
[8] Dietvorst, B. J., Simmons, J. P. and Massey, C.: Over-

coming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them, *Management Science*, Vol. 64, No. 3, pp. 1155–1170 (2018).

[9] Downs, M., Chu, J. L., Yacoby, Y., Doshi-Velez, F. and Pan, W.: CRUDS: Counterfactual Recourse Using Disentangled Subspaces, *Proc. of WHI*, pp. 1–2 (2020).

[10] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G. and Beck, H. P.: The role of trust in automation reliance, *Int. J. of Human-Computer Studies*, Vol. 58, No. 6, pp. 697–718 (2003).

[11] Gemalmaz, M. A. and Yin, M.: Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems, *Proc. of AIES*, pp. 295–306 (2022).

[12] Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B. and Ghosh, J.: Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems (2019).

[13] Kahr, P. K., Rooks, G., Willemsen, M. C. and Snijders, C. C. P.: Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions., *ACM TiiS* (2024).

[14] Kanamori, K., Takagi, T., Kobayashi, K., Ike, Y., Uemura, K. and Arimura, H.: Ordered Counterfactual Explanation by Mixed-Integer Linear Optimization, *Proc of AAAI*, pp. 11564–11574 (2021).

[15] Karimi, A. H., Barthe, G., Schölkopf, B. and Valera, I.: A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations, *ACM CSUR*, Vol. 55, No. 5 (2022).

[16] Karimi, A.-H., Schölkopf, B. and Valera, I.: Algorithmic Recourse: from Counterfactual Explanations to Interventions, *Proc. of FAccT*, pp. 353–362 (2021).

[17] Karimi, A. H., von Kügelgen, J., Schölkopf, B. and Valera, I.: Algorithmic recourse under imperfect causal knowledge: A probabilistic approach, *NeurIPS* (2020).

[18] 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).

[19] Leo, M., Sharma, S. and Maddulety, K.: Machine Learning in Banking Risk Management: A Literature Review, *Risks*, Vol. 7, No. 1, p. 29 (2019).

[20] Lyons, H., Miller, T. and Velloso, E.: Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review, *Proc. of FAccT*, pp. 764–774 (2023).

[21] Lyons, H., Wijenayake, S., Miller, T. and Velloso, E.: What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions, *Proc. of CHI*, pp. 1–15 (2022).

[22] Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M. and Ma, X.: "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making, *Proc. of CHI*, pp. 1–20 (2024).

[23] Mok, L., Nanda, S. and Anderson, A.: People Perceive Algorithmic Assessments as Less Fair and Trustworthy Than Identical Human Assessments, *PACMHCI*, Vol. 7, No. CSCW2, pp. 1–26 (2023).

[24] Mothilal, R. K., Sharma, A. and Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations, *Proc. of FAccT*, pp. 607–617 (2020).

[25] Nourani, M., Honeycutt, D. R., Block, J. E., Roy, C., Rahman, T., Ragan, E. D. and Gogate, V.: Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis, *CHI EA*, pp. 1–8 (2020).

[26] Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E. and Gogate, V.: Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems, *Proc. of IUI*, pp. 340–350 (2021).

[27] Panigutti, C., Beretta, A., Giannotti, F. and Pedreschi, D.: Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems, *Proc. of CHI*, pp. 1–9 (2022).

[28] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W. and Wallach, H.: Manipulating and Measuring Model Interpretability, *Proc. of CHI*, pp. 1–52 (2021).

[29] Rechkemmer, A. and Yin, M.: When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models, *Proc. of CHI*, pp. 1–14 (2022).

[30] Robinette, P., Li, W., Allen, R., Howard, A. M. and Wagner, A. R.: Overtrust of robots in emergency evacuation scenarios, *Proc. of HRI*, pp. 101–108 (2016).

[31] Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A. and Bernstein, A.: Second chance for a first impression? Trust development in intelligent system interaction, *Proc. of UMAP*, pp. 77–87 (2021).

[32] Tominaga, T., Yamashita, N. and Kurashima, T.: Reassessing Evaluation Functions in Algorithmic Recourse: An Empirical Study from a Human-Centered Perspective, *Proc. of IJCAI*, pp. 7913–7921 (2024).

[33] Ustun, B., Spangher, A. and Liu, Y.: Actionable Recourse in Linear Classification, *Proc. of FAccT*, pp. 10–19 (2019).

[34] VanNostrand, P. M., Hofmann, D. M., Ma, L. and Rundensteiner, E. A.: Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding, *Proc. of FAccT*, pp. 1682–1700 (2024).

[35] 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).

[36] Wachter, S., Mittelstadt, B. and Russel, C.: Counterfactual Explanations Without Opening the Black Box: Automated Decisions and thhe GDPR, *Harvard Journal of Law & Technology*, Vol. 20, No. 3, pp. 842–887 (2018).

[37] Wang, R., Harper, F. M. and Zhu, H.: Factors Influencing Perceived Fairness in Algorithmic Decision-Making, *Proc. of CHI*, pp. 1–14 (2020).

[38] Wang, Z. J., Vaughan, J. W., Caruana, R. and Chau, D. H.: GAM Coach: Towards Interactive and User-centered Algorithmic Recourse, *Proc. of CHI*, pp. 1–20 (2023).

[39] Weld, D. S. and Bansal, G.: Intelligible artificial intelligence (2018).

[40] Zhang, Y., Liao, Q. V. and Bellamy, R. K. E.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, *Proc. of FAccT*, pp. 295–305 (2020).

[41] 東京都政策企画局：自動車利用と環境に関する世論調査，`https://www.metro.tokyo.lg.jp/tosei/hodohappyo/press/2023/03/29/documents/01_full.pdf` (2023).