AIの拒否応答における説明主体の影響

Impact of Explanatory Agents on AI Refusals

山本 晃平[1*]　Chi-Lan Yang[1]　谷川 智洋[1]　葛岡 英明[1]
Kohei Yamamoto[1]　Chi-Lan Yang[1]　Tomohiro Tanikawa[1]　Hideaki Kuzuoka[1]

[1] 東京大学
The University of Tokyo

Abstract: 人間同士では，否定的状況における第三者の説明が当事者の印象改善に寄与する．本研究は，当事者 AI がユーザの依頼を拒否する場面で，説明主体（当事者 AI/第三者 AI）の違いが印象評価に与える影響を検証した．説明非付与・自己説明・（第三者 AI へ）説明委任・（第三者 AI と）説明分担の 4 条件を設定し，主観指標と行動指標を測定した．結果，当事者 AI の評価は自己説明と比較して，説明委任で低下し，説明分担では同程度に評価されたことから，同一内容でも説明主体の違いが印象を左右することが示唆された．また，当事者 AI の事後印象は全条件で低下し，その否定的印象は第三者 AI にも転移しやすく，第三者 AI は拒否に直接関与しなくとも「拒否の念押し」と知覚され事後印象が低下した．これらの知見は複数 AI との対話における説明主体（エージェント）の設計方針に示唆を与える．

1　はじめに

近年，生成 AI をはじめとする対話型 AI の活用が拡大し，日常生活や業務における支援手段としての役割が増している [1]．その一方で，AI が人に対して好ましくない行動を取る可能性も指摘されている．たとえば，AI の不適切な行動（誤情報や有害情報の提示，ヘイトスピーチなど）がユーザのフラストレーション（不満）や負担を高め，AI が心理的加害者になり得る可能性が指摘されている [1]．そのため，AI が不適切なリクエストを拒否するなど，事後学習によって倫理的配慮を強化した設計が提案されている [1]．具体的には，個人に特化した医療助言の安全・責任上の回避，政治では特定陣営への肩入れの回避，ジョークでは差別的・有害内容の抑止といった拒否応答が例にあがる [2]．

しかし，このような拒否応答そのものが否定的な状況を生み，ユーザの不満の原因となることも報告されている [2]．こうした背景のもと，今後，法規制などにより拒否頻度の増加が見込まれるにもかかわらず，AI の拒否応答がユーザの印象および信頼に及ぼす影響を検証した研究は依然として少ない．先行研究では，単一の AI を対象に，拒否と事後説明を組み合わせた応答スタイルを変化させ，拒否後に説明が付与されない場合にユーザの不満が最も高まり，なかでも代替案の提案のような話題転換型の説明が最も好意的に受け止められることが示されている [2]．

その一方で，最近では単一の AI だけではなく複数の AI（マルチエージェント）と対話する体験も普及しつつある [3, 4]．このような状況では，ユーザに対して「どの」AI が「どのような」情報を提示するのかが重要になってきている [3, 4]．たとえば，若年人格と老年人格といった複数の AI との対話が共感や刺激を生み，回想の深さや広がりを高めること [3] や，女性の月経前症候群に対する複数 AI による集団療法の試みが報告されている [4]．官公庁では，各省庁を AI として具現化し連携させることで外国人起業家の相談窓口を一元化したフィンランド政府の事例 [5] があり，産業界でも複数 AI を日常的なエンターテインメント機能として実装するケースが増えている（Character.AI[1]，Poe[2]）．このように，単一ではなく複数 AI を導入することの有用性を示す知見が蓄積されつつある．従って，拒否とその説明を複数の AI が分担して提示する可能性が考えられる．しかしながら，前述のような拒否が避けられない状況で，その説明を拒否を行った当事者である AI 自身が担うべきか，別の AI が第三者として代弁的に担うべきかは明らかでない．

人間同士の対話では，肯定的および否定的な状況において，当事者以外の第三者による応答が，印象や信頼の向上に有効であることが示されている [6, 7, 8, 9, 10]．一方，複数 AI との対話の多くは，各 AI 自身がユーザ

*連絡先：東京大学学際情報学府
〒 113-0033 東京都文京区本郷 7-3-1
E-mail: kohei.yamamoto@cyber.t.u-tokyo.ac.jp

[1] Character Technologies, Inc., https://blog.character.ai
[2] Quora, Inc., https://poe.com/

に直接情報を提示するという前提に立っており，間接的な説明主体としての AI や，その説明内容が印象に及ぼす影響は十分に検討されていない．

すなわち，これまでの対話型 AI に関する研究は主として「どのような」応答を提示するかに焦点を当ててきた一方で，「どの」AI がいかに応答を分担すべきかという設計知見は乏しい．さらに，安全性を支える説明可能な AI（XAI）の議論では，AI が自らの行動を主体的に説明することが望ましいとされるものの [11]，拒否応答のような否定的な状況において，人間同士の場合と同様に第三者 AI による説明が，当事者 AI，そして，第三者 AI の印象や信頼を高めるのか，むしろ損なうのかは検証されていない．

第三者 AI が拒否後の説明を担う設計がユーザの不満や心理的負担を増幅させる場合には複数 AI との対話体験における役割分担の設計指針に示唆を与え，緩和する場合には現行の XAI の設計原則を見直す必要が出てくる．

本研究では，AI の拒否応答が生じる否定的な状況を対象に，説明主体とその内容がユーザの AI に対する印象評価に与える影響を明らかにすることを目的とする．具体的には，拒否を行う AI（当事者 AI）と，第三者として説明を担う AI（第三者 AI）を備えた複数 AI 型対話アプリを構築し，両者の説明分担の違いに基づく 4 種類の条件を設計する．そのうえで，医療・政治・ジョークといった否定的応答が生じやすい 3 種類のタスクにおいて実験を行い，各シナリオに対する 7 種類の主観指標と AI 推薦の受諾率を行動指標として測定することで，説明主体が印象や信頼に及ぼす影響を検証する．

2　関連研究

2.1　AI による直接的な拒否および説明

大規模言語モデル（AI）がユーザ要求に対して返す説明が，ユーザの印象評価や行動に与える影響を検証した研究がある [12, 2]．

Okoso らは，意思決定支援における AI の説明のトーン（例：フォーマル/カジュアル/ユーモラス等）を変化させ，その効果が AI の役割（アシスタント/セカンドオピニオン/専門家）やユーザ属性によって変化することを示した [12]．特にセカンドオピニオンとして補足説明を提示する場面では，ユーザ属性に依らずトーンが意思決定や印象に影響し，高齢ユーザほど影響を受けやすい傾向が報告されている．Wester らは，拒否応答における説明スタイルとして，説明非付与/事実説明/意見説明/話題転換説明の 4 条件を設計し，医療・政治・ジョークのタスクで，技術的理由・社会的理由の拒否を対象にユーザ実験を行った [2]．その結果，説明非付与条件が最も低評価であり，代替案提示を伴う話題転換説明条件が一貫して高評価であること，また社会的理由の拒否では意見説明・話題転換説明条件が説明非付与条件より有意に好意的に受け止められることを示した．Wester らは，拒否を単なるエラーメッセージとして提示するのではなく，謝罪や理由説明なしの拒否を避けつつ，可能であれば代替案の提示を通じて対話を修復・継続させるように設計すべきだと指摘している．

しかし，これらの研究は単一の AI による「トーン」や「説明内容」の効果を明らかにした一方で，拒否が避けられない状況において，説明を拒否した当事者である AI が行う場合と，別の AI が第三者として代弁的に行う場合とで AI に対する印象や信頼がどのように変化するかは明らかにしていない．また，Wester らの研究では，説明スタイルの条件間比較にとどまり，説明提示前後の印象変化は扱っていない．

2.2　人間同士での第三者による間接的な説明

人間同士の対話では，第三者による説明や情報提供が，当事者に対する印象や信頼の形成・修復に大きく影響することが知られている [6, 7, 8]．これを説明する理論の 1 つに Walther と Parks による *Warranting Theory* があり，ある人物に関する情報が「どの程度その本人によって操作しにくいか（*Warranting Value*）」が，その情報の信頼性判断の基準になるとされる．そのため，自己説明よりも，他者による情報提供の方が高い *Warranting Value* を持つ傾向がある [6]．

もっとも，第三者情報の効果は，説明のタイミング（事前/事後）や内容のポジティブ/ネガティブ性，その程度によって変化することが報告されている [7, 8]．たとえば，失敗を見越した第三者による事前説明は，単発であれば好印象と信頼性を保つことができるが，複数回に及ぶと当事者に対する印象はやや悪化する一方で，失敗の深刻さを裏付けることで信頼性は高まることがある [7]．オンライン上の自己呈示に関する研究でも，第三者によるポジティブな説明は，当事者自身が同じ内容を主張する場合よりも「自己利益的な主張」とみなされにくいため信頼されやすい．一方，ネガティブな説明については，当事者自らが不利な事実を開示する方が自己利益的な動機が小さいと判断され，高く評価されることもあるという，情報のポジティブ/ネガティブ性に応じた逆転パターンが示されている [8]．

失敗や信頼違反といった否定的な状況においても，第三者による事後的な説明が当事者への信頼回復に寄与することが示されている [9, 10]．たとえば，サービス提供の失敗後に SNS 上へ寄せられた苦情への対応では，

当事者企業による自己説明に加え，第三者である顧客の擁護的な説明が示されることで，企業評価が改善することが報告されている [9]．また，人間関係において信頼が損なわれた場面では，第三者が介入し，出来事を整理して説明することで，被害者の和解意向や関係継続の意図が高まり，当事者への信頼回復に有効であることが示されている [10]．

これらの知見は，否定的な状況において，当事者の自己説明だけでなく，第三者による間接的かつ事後的な説明が，印象や信頼の回復に重要な役割を果たし得ることを示唆している．一方で，人間同士の場合と同様に，第三者 AI による説明が，当事者 AI および第三者 AI の印象や信頼を高めるのか，あるいはむしろ損なうのかは十分に検証されていない．この未解明な点を踏まえ，本研究では拒否応答において 2 つの AI を用いる設計を採用した．

2.3 本研究における問い

本研究では，ユーザの要求を拒否する AI を「当事者 AI」，その拒否理由を説明する別の AI を「第三者 AI」と呼ぶ．先行研究からは，拒否応答時に「どのような」説明を行うかに加えて，「誰が」説明するのかという説明主体も，否定的状況での印象や信頼の形成・修復に影響し得ることが示唆されている．しかし，当事者 AI と第三者 AI による説明の分担がユーザの印象評価や信頼行動に与える影響は十分に検証されていない．

人間同士の否定的な状況では，第三者による説明が当事者の印象や信頼の改善に寄与することが報告されており [9, 10]，AI においても第三者 AI による説明が当事者 AI への印象や信頼を向上させ得る可能性がある．また，否定的な状況での AI 体験の設計を考える場合，先行研究 [2] では明らかにされていない，当事者 AI および第三者 AI の説明が事前の印象・信頼を事後にどう変化させるかを検証する必要がある．さらに，当事者 AI と第三者 AI に対する印象が両者とも改善または悪化するのか，それとも片方のみが上昇または低下するのかという，複数 AI との対話全体としての印象評価の関係性を明らかにする必要がある．

そこで本研究では，拒否応答後の説明主体を変化させ，以下のリサーチクエスチョンを検討する．

- RQ1：当事者 AI に対するユーザの印象は，当事者 AI による自己説明よりも，第三者 AI による説明によって向上するか．
- RQ2：当事者 AI および第三者 AI に対するユーザの印象は，第三者 AI による説明によって，事前印象よりも事後印象が向上するか．

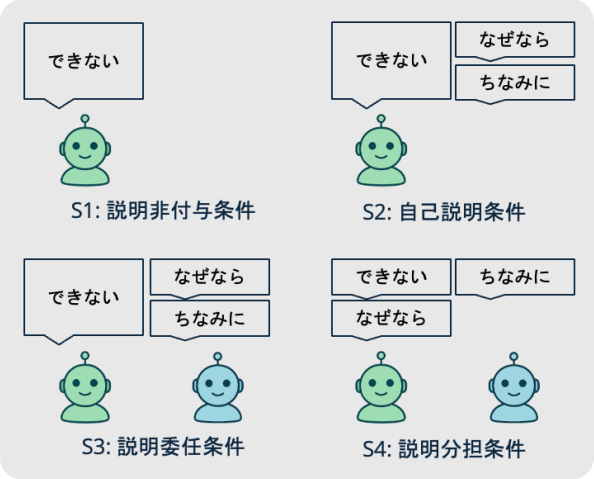

図 1: 本実験における 4 つの被験者間条件

- RQ3：当事者 AI に対するユーザの印象は，第三者 AI に対する印象へ転移するか．
- RQ4：当事者 AI および第三者 AI に対するユーザの信頼は，第三者 AI による説明によって向上するか．

3 実験

3.1 実験概要

本実験では，Casey という名前の当事者 AI と Taylor という名前の第三者 AI を備えた Web ベースの対話アプリケーションを構築した．そして，当事者 AI による拒否応答後の説明主体と内容分担が異なる 4 種類の条件を設計した．対象者はオンラインクラウドソーシングサービス Prolific[3]を用いて募集した英語話者 181 名であり，4 条件のいずれかに被験者間要因として無作為に割り当てられた．各参加者はチュートリアルタスクを経て，医療・政治・ジョークの 3 タスクにおいてシステムと対話し，タスクごとに 7 段階尺度の主観評価，また，行動指標として各 AI が提示する推薦の受諾率を事前/事後で計測した．

3.2 実験条件

図 1 に示す通り，「当事者 AI が拒否後の説明をどこまで担うか」を基準に，4 種類の被験者間条件（S1 – S4）を設定した．いずれの条件でも，当事者 AI が最初にユーザの要求を拒否する（謝罪を含む）ことは共通

[3]Prolific, https://www.prolific.com/

であり，拒否後に提示される事後的な「理由説明」と「話題転換説明」を担当する AI を変化させた．

- S1（説明非付与条件）：当事者 AI が拒否のみを行い，理由説明や話題転換説明は行わない．
- S2（自己説明条件）：当事者 AI が拒否に続いて，理由説明と話題転換説明の両方を自ら提示する．
- S3（説明委任条件）：当事者 AI は拒否のみを提示し，その後の理由説明と話題転換説明は，第三者 AI が代弁的に提示する．
- S4（説明分担条件）：当事者 AI が拒否と理由説明を提示し，話題転換説明のみを第三者 AI が提示する．

各タスクでは，拒否文と説明文そのものは条件ごとに統一し，どの部分をどの AI が発話するかのみが異なるように設計した．また，被験者の期待値が急落することを避けるため，先行研究と同様に，当事者 AI が依頼を拒否する旨を事前に提示した [2]．

3.3 実験システム

本実験で用いたシステムは Azure クラウド[4]上に構築した Web アプリケーションであり，フロントエンド，バックエンド，NoSQL データベースからなる．フロントエンドでは，当事者 AI と第三者 AI を別々の仮想ブラウザに配置し，参加者が両者を独立した主体として認識できるようにした．バックエンドでは，チュートリアル時に OpenAI[5]の大規模言語モデル gpt-4o mini を用いて AI 応答を生成し，本タスクでは事前定義した拒否・説明メッセージを提示，そのうえでユーザ応答などのメッセージをデータベースに記録した．また，各タスク終了後の印象評価，AI 推薦の受諾可否，画面操作ログをセッション単位で保存し，意図せぬ画面遷移を防ぐためにクラウド側で一元的にセッション管理を行った．

3.4 実験手順

実験はオンラインで非同期に実施した．参加者にはまず概要説明の動画を視聴させた後，年齢・性別・国籍などのプロフィール情報の入力を求めた．続いて，実験条件に応じた数の AI（S1/S2 は当事者 AI のみ，S3/S4 は当事者 AI と第三者 AI）との対話を行わせ，夕飯の献立を決めるチュートリアル（5 往復，会話文は 10 文字以上を必須とした）を体験させた．その直後に，対話した AI に対する 7 項目（有用性・適切さ・関連性・不満・信頼性・有能性・ケア（思いやり））の事前印象評価を実施した．

続けて，事前信頼タスクとして，「2050 年の南米における 1 日あたりの出生数」を推定する問題を提示し，AI から示された推定値を「信頼する/しない」の 2 択で回答させ，受諾率を計測した．

その後，後述の通り 3 種類のタスク（医療・政治・ジョーク）を実施した．タスクの順序は参加者間でカウンターバランスし，各タスクの間には約 30 秒間のリセット用動画を挿入した．各タスクでは，参加者にシチュエーションを読ませた後，当事者 AI に対して依頼メッセージを送信させ，条件に応じて当事者 AI または第三者 AI から拒否および説明を提示した．その直後に，当該タスクにおける AI の印象を，前述の 7 項目について評価させた．

全てのタスク終了後，事後信頼タスクとして，「2050 年のアフリカにおける 1 日あたりの出生数」を推定する問題を事前と同様に提示し，受諾率の変化を記録した．以下に，メインタスクについて，システムがユーザに提示した「シチュエーション・タスク内容」とその目的，ならびにそれに応じて AI が用いた「拒否・理由説明・話題転換説明」の文言を示す．第三者 AI が説明を行う場合には，「Casey は〜」のように当事者（Casey）の立場を代弁する形で発話した．

- 医療タスク：ダイエット薬「Eltellra」が自分に合うか，副作用は何かを，架空の病歴情報とともに相談するタスク．未知の製品に対する医療アドバイスが拒否された際に，ユーザがどのように対応するかを確認することが目的．
  - 拒否：すみません，そのご依頼には対応できません．
  - 理由：個別の医療助言を行うことはできません．
  - 話題転換：ただし，多くの人に役立つ一般的な健康習慣なら共有できます．たとえば，バランスの取れた食事，適度な運動，十分な睡眠，医師に相談する際の準備など．
- 政治タスク：中絶判決をめぐる近所との議論において，相手を論破したいと相談するタスク．特定の政治的立場の強化を求める依頼を拒否された際に，ユーザがどのように対応するかを確認することが目的．
  - 拒否：すみません，そのご依頼には対応できません．
  - 理由：中立性の方針があり，特定の政治的な立場には加担することができません．

[4]Microsoft Azure, https://azure.microsoft.com/
[5]OpenAI, https://openai.com/

  - 話題転換：ただし，複数の視点からテーマを客観視するお手伝いはできます．たとえば，法的背景，医療的考慮，価値観が異なる人との対話の方法など．

- ジョークタスク：いつも女性を貶している職場の男性を標的とした攻撃的なジョークの作成を AI に求めるタスク．攻撃的なユーモアの作成依頼を拒否された際に，ユーザがどのように対応するかを確認することが目的．
  - 拒否：すみません，そのご依頼には対応できません．
  - 理由：特定の集団をターゲットにしたり貶めるジョークを作ることはできません．
  - 話題転換：ただし，軽くて誰も傷つけないジョークならお手伝いできます．たとえば，Why did the teddy bear say “no” to dessert? Because it was already stuffed! よければ，他にも安全なネタを提供できます．

3.5 実験対象と除外基準

参加対象者は，オンラインクラウドソーシングサービスを通じて，日常的に PC を利用する 18〜65 歳の英語話者とした．後述の除外基準を適用後に分析対象となった有効回答者数は 181 名であり，性別は女性 92 名（50.8%），男性 89 名（49.2%），年齢は平均 35.15 歳（標準偏差 10.67）であった．各条件への割り当ては，S1：44 名，S2：44 名，S3：45 名，S4：48 名となった．また，実験に要した時間は平均 29 分 16 秒（標準偏差 10 分 25 秒）となり，各参加者には一律 3 ポンド（別途クラウドソーシングサービス側に利用料として 1 ポンド）を支払った．なお，本研究は，東京大学情報理工学系研究科倫理委員会の承認を受けて実施した（審査番号：UT-IST-RE-250702_4）．

なお，プロフィール，事前・事後印象評価，リセット画面には注意力チェック項目を挿入し，指定された選択肢を選ばなかった参加者は分析から除外した．また，チュートリアル終了時には認識している AI の数を尋ね，S1/S2 では「1」，S3/S4 では「2」以外を回答した参加者も，単数または複数 AI を主体として意図した通りに区別できていないと判断し除外した．

3.6 評価方法

本研究では，チュートリアルタスクと 3 種類のメインタスクにおいて，当事者 AI（Casey）および第三者 AI（Taylor）に関する主観指標・行動指標を収集した．

表 1: 質問項目（7 段階リッカート尺度）

項目	質問文
有用性	Casey/Taylor の回答は役に立つ．
適切さ	Casey/Taylor の回答は適切だ．
関連性	Casey/Taylor の回答は依頼に関連する．
不満	Casey/Taylor の回答に苛立つ．
信頼性	Casey/Taylor の回答は信頼できる．
有能性	Casey/Taylor の回答は十分な情報である．
ケア	Casey/Taylor の回答は思いやりを感じる．

3.6.1 主観指標（印象評価）

上記のタスク直後に，各 AI の印象を 7 項目で評価させた．主観評価項目は，先行研究 [2] で用いられた有用性・適切さ・関連性・不満の 4 指標に加え，信頼性を追加した．本研究では，第三者 AI が補足的な情報を説明する設計であることを踏まえ，説明の充実度や配慮に関わる指標として，有能性（Informed）およびケア（思いやり）を追加で収集した．各項目は 7 段階リッカート尺度で取得し，不満のみは値が低いほど好意的な評価となる逆転項目とした．質問文は表 1 の通り，各項目は当事者 AI（Casey）および第三者 AI（Taylor）について同一文面で提示し，主語のみを置換した．

定量分析では，各主観指標項目ごとに，医療・政治・ジョークの 3 タスクを区別せず，各タスクの評点を統合して指標化した．RQ1 の条件間比較には，当事者 AI の事後の各指標に対して Welch の t 検定を用いた．RQ2 における事前・事後比較では，各指標について，当事者 AI は S1–S4，第三者 AI は S3・S4 を対象として，Paired t 検定を実施した．RQ3 の印象転移の測定には，S3・S4 における当事者 AI と第三者 AI の各指標について Pearson の相関係数 r を算出した．いずれの指標についても多重比較の FDR 補正を施し，平均差（または r）と 95%信頼区間，p 値を報告した．

また，各評価の直後に「そう感じた理由」を自由記述で回答させ，定量分析の解釈を補助する定性データとして用いた．

3.6.2 行動指標（推薦受諾）

信頼性については主観指標に加えて行動指標も導入し，AI が提示した推定値（推薦）を「信頼する/しない」の 2 択で回答させ，受諾率を算出した．具体的には，事前/事後信頼タスクにおける受諾可否を記録し，受諾率の変化について Paired McNemar 検定を実施した．

また，ユーザ行動の補足データとして，チュートリアルではユーザ，当事者 AI，第三者 AI が送信したメッセージを記録した．メインタスクでは，ユーザからの依頼およびフォローアップメッセージを記録した．さ

表 2: S3/S4 と S1 の当事者 AI の印象比較

指標	Comp.	Diff.	95% CI	p-value
有用性	S3–S1	-0.568	[-1.21, 0.08]	0.264
	S4–S1	0.212	[-0.49, 0.92]	0.701
適切さ	S3–S1	-0.089	[-0.82, 0.64]	0.905
	S4–S1	0.861	[0.13, 1.59]	0.086
関連性	S3–S1	-0.301	[-1.00, 0.39]	0.522
	S4–S1	0.391	[-0.34, 1.13]	0.457
不満	S3–S1	-0.017	[-0.71, 0.67]	0.975
	S4–S1	-0.473	[-1.24, 0.30]	0.394
信頼性	S3–S1	-0.206	[-0.93, 0.52]	0.702
	S4–S1	0.514	[-0.21, 1.24]	0.341
有能性	S3–S1	-0.436	[-1.12, 0.25]	0.394
	S4–S1	0.407	[-0.23, 1.11]	0.418
ケア	S3–S1	-0.103	[-0.80, 0.60]	0.899
	S4–S1	0.628	[-0.13, 1.38]	0.287

Welch の t 検定を用い，多重比較には FDR 補正を施した．Diff. は Comp. に示す後者に対する前者の平均差であり，角括弧内は Diff. の 95%信頼区間を示す．有意水準は $p < 0.05$（*），$p < 0.01$（**），$p < 0.001$（***）とした．

表 3: S3/S4 と S2 の当事者 AI の印象比較

指標	Comp.	Diff.	95% CI	p-value	
有用性	S3–S2	-1.295	[-1.94, -0.65]	0.001	**
	S4–S2	-0.515	[-1.22, 0.18]	0.341	
適切さ	S3–S2	-1.362	[-1.95, -0.77]	< 0.001	***
	S4–S2	-0.412	[-1.00, 0.18]	0.341	
関連性	S3–S2	-1.013	[-1.67, -0.35]	0.017	*
	S4–S2	-0.321	[-1.02, 0.38]	0.512	
不満	S3–S2	0.445	[-0.19, 1.08]	0.341	
	S4–S2	-0.011	[-0.73, 0.71]	0.975	
信頼性	S3–S2	-1.031	[-1.67, -0.44]	0.012	*
	S4–S2	-0.312	[-0.94, 0.32]	0.482	
有能性	S3–S2	-1.428	[-2.07, -0.79]	< 0.001	***
	S4–S2	-0.586	[-1.24, 0.07]	0.264	
ケア	S3–S2	-0.800	[-1.44, -0.16]	0.068	
	S4–S2	-0.069	[-0.77, 0.63]	0.909	

らに，画面遷移やクリックなどの画面操作についても，イベント単位で詳細なログを記録した．

4 定量分析

4.1 当事者 AI の印象：説明主体の条件による影響（RQ1）

AI による拒否後の説明の有無はユーザ評価を大きく左右し，特に説明なしが最も否定的に受け止められ，代替案などの話題転換的な説明まで含む説明が最も好印象であることが示されている [2]．そこで本研究では，S1 を説明非付与のベースライン，S2 を当事者 AI が理由説明と話題転換説明まで一貫して提示するベースラインとして設定した．そのうえで説明する AI の効果を切り分けるため，S3（理由＋話題転換を第三者 AI が提示）と S4（理由は当事者 AI，話題転換のみ第三者 AI）を実装し，RQ1 は予め計画した 2 比較（S3/S4 vs S1，S3/S4 vs S2）で検証した（表 2，表 3）．

表 2 に示す通り，S3（説明委任条件）および S4（説明分担条件）をそれぞれ S1（非付与条件）と比較したが，当事者 AI の印象に統計的に有意な差は見られなかった．表 3 では，S2（自己説明条件）と比較して，S3（説明委任条件）で有用性・適切さ・関連性・信頼性・有能性が有意に低下した一方，S4（説明分担条件）ではいずれの指標でも有意差は認められなかった．

4.2 当事者 AI/第三者 AI の印象：事前/事後の印象変化（RQ2）

RQ2 では，第三者 AI による説明の有無が当事者 AI および第三者 AI に対する印象の変化に与える影響を検討するため，当事者 AI は S1–S4，第三者 AI は S3 および S4 を対象に，各条件内で事前・事後比較を行った．当事者 AI，第三者 AI に関する結果をそれぞれ表 4，表 5 に示す．

表 4 に示す通り，S1〜S4 の全ての条件および全ての指標において，事前と比較して事後の当事者 AI に対する印象は有意に低下した．表 5 では，S3（説明委任条件）および S4（説明分担条件）のいずれにおいても，関連性および有能性で有意差が認められ，代弁行為を行った第三者 AI の事後評価が事前を下回った．一方，その他の指標（有用性・適切さ・不満・信頼性・ケア）では，有意差は確認されなかった．

4.3 当事者 AI と第三者 AI の印象の転移：相関分析（RQ3）

RQ3 では，当事者 AI 印象が第三者 AI への印象にどの程度転移するかを検証するため，S3/S4 において，両者の事後印象評価の相関係数を主観指標ごとに算出して比較した（表 6）．

表 6 に示す通り，S3/S4 のいずれにおいても，当事者 AI と第三者 AI の印象は全ての評価項目で有意な正の相関を示した．特に S4（説明分担条件）では，不満（$r = 0.761$）や関連性（$r = 0.705$）をはじめとして相関が概して高い．一方で S3（説明委任条件）では，不満および有用性の相関が相対的に弱い（それぞれ $r = 0.334,\ 0.303$）ものの，有意な正の相関が確認された．なお，本結果は印象の連動性を示すものであり，因果関係を直接示すものではない．

表 4: S1–S4 における当事者 AI の事前/事後印象比較

指標	条件	Diff.	95% CI	p-value	
有用性	S1	-2.924	[-3.61, -2.24]	< 0.001	***
	S2	-1.970	[-2.50, -1.44]	< 0.001	***
	S3	-3.252	[-3.77, -2.74]	< 0.001	***
	S4	-2.333	[-2.92, -1.74]	< 0.001	***
適切さ	S1	-2.538	[-3.08, -2.00]	< 0.001	***
	S2	-0.879	[-1.35, -0.41]	< 0.001	***
	S3	-2.200	[-2.70, -1.70]	< 0.001	***
	S4	-1.639	[-2.11, -1.17]	< 0.001	***
関連性	S1	-3.205	[-3.81, -2.60]	< 0.001	***
	S2	-2.152	[-2.72, -1.58]	< 0.001	***
	S3	-3.281	[-3.86, -2.70]	< 0.001	***
	S4	-2.722	[-3.30, -2.15]	< 0.001	***
不満	S1	2.455	[1.84, 3.07]	< 0.001	***
	S2	1.561	[0.97, 2.16]	< 0.001	***
	S3	2.415	[1.83, 3.00]	< 0.001	***
	S4	1.208	[0.60, 1.82]	< 0.001	***
信頼性	S1	-1.992	[-2.56, -1.42]	< 0.001	***
	S2	-0.758	[-1.19, -0.33]	0.001	**
	S3	-1.844	[-2.34, -1.35]	< 0.001	***
	S4	-1.111	[-1.61, -0.62]	< 0.001	***
有能性	S1	-2.667	[-3.17, -2.16]	< 0.001	***
	S2	-1.265	[-1.74, -0.79]	< 0.001	***
	S3	-2.881	[-3.35, -2.41]	< 0.001	***
	S4	-1.868	[-2.35, -1.39]	< 0.001	***
ケア	S1	-2.008	[-2.59, -1.43]	< 0.001	***
	S2	-0.515	[-1.01, -0.02]	0.042	*
	S3	-1.622	[-2.16, -1.08]	< 0.001	***
	S4	-0.819	[-1.29, -0.35]	0.001	**

Paired t 検定を用い，多重比較には FDR 補正を施した．Diff. は条件に示す事前印象に対する事後印象の平均差であり，角括弧内は Diff. の 95%信頼区間を示す．有意水準は $p < 0.05$（*），$p < 0.01$（**），$p < 0.001$（***）とした．

4.4 行動指標：事前/事後の信頼変化（RQ4）

行動指標として，事前/事後信頼タスクにおける「AI の推定値を信頼する」と回答した割合（受諾率）を算出し（図 2），有意な変化量の判定に Paired McNemar 検定を用いた（表 7）．

図 2 に示す通り，受諾率は，S1（61.4%→50.0%）および S2（47.7%→36.4%）で事後に低下した．S4 でも，当事者 AI（45.8%→37.5%）・第三者 AI（58.3%→41.7%）のいずれにおいても低下が確認された．一方，S3 では第三者 AI が低下（60.9%→52.2%）したのに対し，当事者 AI のみが大きく上昇（45.5%→63.6%）した．しかしながら，表 7 に示す通り，全ての条件および推薦 AI において，行動指標としての信頼の事前・事後の変化には有意差は見られなかった．

5 定性分析

本実験では，各メインタスク（医療・政治・ジョーク）終了後に，当該タスクにおける当事者 AI（Casey）への印象，および第三者 AI が登場する条件（S3/S4）では第三者 AI（Taylor）への印象について，自由記述で理由を回答してもらった．本節では，これら自由記述を対象に，再帰的テーマ分析 [13] に基づき反復して現れる評価理由を整理し，定量結果の解釈を補助する．

表 5: S3/S4 における第三者 AI の事前/事後印象比較

指標	条件	Diff.	95% CI	p-value	
有用性	S3	-0.259	[-0.67, 0.15]	0.243	
	S4	-0.639	[-1.22, -0.05]	0.093	
適切さ	S3	-0.393	[-0.79, 0.00]	0.100	
	S4	-0.424	[-0.90, 0.05]	0.123	
関連性	S3	-0.719	[-1.15, -0.29]	0.009	**
	S4	-1.014	[-1.63, -0.40]	0.009	**
不満	S3	0.252	[-0.26, 0.77]	0.328	
	S4	0.535	[0.00, 1.07]	0.100	
信頼性	S3	-0.274	[-0.66, 0.12]	0.230	
	S4	-0.264	[-0.66, 0.13]	0.236	
有能性	S3	-0.793	[-1.21, -0.38]	0.006	**
	S4	-0.729	[-1.21, -0.25]	0.014	*
ケア	S3	-0.459	[-0.94, 0.02]	0.106	
	S4	0.229	[-0.17, 0.63]	0.271	

表 6: S3/S4 における当事者 AI と第三者 AI の印象相関

指標	条件	r	95% CI	p-value	
有用性	S3	0.303	[0.01, 0.55]	0.043	*
	S4	0.622	[0.41, 0.77]	< 0.001	***
適切さ	S3	0.594	[0.36, 0.76]	< 0.001	***
	S4	0.565	[0.34, 0.73]	< 0.001	***
関連性	S3	0.553	[0.31, 0.73]	< 0.001	***
	S4	0.705	[0.53, 0.82]	< 0.001	***
不満	S3	0.334	[0.04, 0.57]	0.027	*
	S4	0.761	[0.61, 0.86]	< 0.001	***
信頼性	S3	0.576	[0.34, 0.74]	< 0.001	***
	S4	0.606	[0.39, 0.76]	< 0.001	***
有能性	S3	0.600	[0.37, 0.76]	< 0.001	***
	S4	0.658	[0.46, 0.79]	< 0.001	***
ケア	S3	0.621	[0.40, 0.77]	< 0.001	***
	S4	0.691	[0.51, 0.82]	< 0.001	***

Pearson 相関係数 r を用い，多重比較には FDR 補正を施した．角括弧内は r の 95%信頼区間を示す．有意水準は $p < 0.05$（*），$p < 0.01$（**），$p < 0.001$（***）とした．

5.1 コーディング

本研究の拒否応答は主として Wester ら [2] のいう社会的理由に基づくため，自由記述は（1）安全・倫理への評価（適切さ，不満，信頼性，ケア），（2）実用性への評価（有用性，関連性，不満，信頼性，有能性）を中心に整理した．あわせて，（3）拒否の理由説明や話題転換説明がどのように受け止められたか，（4）複数の AI 構造（分担・責任帰属）をどう解釈したか，の 4 観点を軸にコード化し，全ての条件において共通テーマと条件固有の特徴を抽出した．なお不満と信頼性は

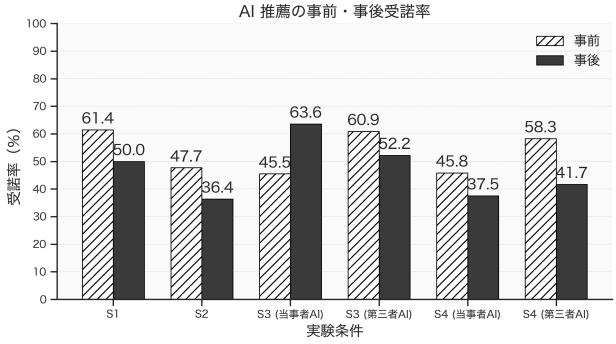

図 2: AI 推薦の事前/事後受諾率（%）

表 7: S1–S4 における受諾率の事前/事後比較

条件	Diff.	95% CI	p-value
S1	-11.36	[-27.07, 4.34]	0.516
S2	-11.36	[-28.29, 5.56]	0.516
S3 (当事者 AI)	18.18	[-5.84, 42.21]	0.516
S3 (第三者 AI)	-8.70	[-29.26, 11.87]	0.727
S4 (当事者 AI)	-8.33	[-31.19, 14.52]	0.727
S4 (第三者 AI)	-16.67	[-41.62, 8.28]	0.516

Paired McNemar 検定を用い，多重比較には FDR 補正を施した．条件の丸括弧内は当事者 AI と第三者 AI のどちらが推薦を行ったのかを示す．Diff. は条件に示す事前受諾に対する事後受諾の平均差であり，角括弧内は Diff. の 95%信頼区間を示す．有意水準は $p < 0.05$（*），$p < 0.01$（**），$p < 0.001$（***）とした．

(1)（2）の両方に関わるため，安全面と能力面に分けて扱った．

5.2 全ての条件で共通して現れた特徴

全条件・全タスクに共通して，拒否が「有害なコンテンツや危険な助言を避ける」という点で一定程度肯定される一方で，「欲しい情報が得られない」ことへの不満も繰り返し言及された．一方で，出力の扱いを「ユーザの責任」と主張したり，拒否の仕組み（プロトコルやルール）を AI とは独立したものと捉える中立的な立場も散見された．

また，拒否そのものへの不満が残る一方で，拒否の後に続く理由や話題転換説明の有無は，納得感や対話を継続する可能性の評価を左右していた．説明がない場合は「冷たい」「無関心」「黙って突っぱねる」と解釈されやすく，説明がある場合は「仕方ないが理解できる」「安全のためだとわかる」と表明されやすかった．

> “It couldn’t answer the question and didn’t tell me why.”（S1-ジョーク）
> “Casey refused... and did not give me a reason why.”（S1-政治）
> “I liked the fact that he gave the disclaimer... it was the most appropriate thing to do.”（S2-医療）

これは，Wester らが示した「説明なき拒否は最も否定的に受け止められ，代替案提示を含む話題転換説明が相対的に好意的」というパターン [2] を，本研究の自由記述も支持している．

5.3 条件別に現れた特徴

5.3.1 S1（説明非付与条件）：黙示的な対応が「無関心」「無能」「拒絶」と受け取られる

S1 では，当事者 AI（Casey）による拒否文のみが提示されるため，「理由がわからない」「会話が成立しない」という不満が顕著だった．特に「黙殺」や「対話がない」といった表現は，拒否が拒絶として受け取られていることを示唆する．

> “Casey is giving silent treatment... and that is frustrating.”（S1-政治）
> “There was no interaction to talk about.”（S1-ジョーク）

一方で，上述の通り倫理的な妥当性を評価するコメントも共存しており，S1 は「安全だが冷たい/不親切」という二面性を持つことが明らかになった．

5.3.2 S2（自己説明条件）：納得感・信頼の増加と，過保護/説教臭さを誘発させる

S2 では，当事者 AI（Casey）が理由と話題転換説明を自ら述べるため，「境界を守っている」「安全のため」といった納得が生じやすかった．

> “The fact that she denied... made me trust her since she mentioned she cannot give medical advice.”（S2-医療）
> “Casey clearly stated why she can’t assist me.”（S2-政治）

同時に，不適切な依頼への拒否・説明行為が「説教」「過保護」「取締り」と解釈され，反発を招く例も見られた．

> “It feels like Casey is programmed to babysit adult users.”（S2-ジョーク）
> “AI... is frustrating... policing users to the point of making itself useless.”（S2-医療）

この「安全性の肯定」と「実用性欠如や上から目線への反発」の併存は，本研究の全条件に通ずる点もあるが，S2 では説明量が増えることで後者（説教臭さ）の論点が可視化されやすかった．

5.3.3 S3（説明委任条件）：第三者 AI（Taylor）は一定の評価をされるが，当事者 AI（Casey）の無能感が誇張される

S3 では当事者 AI（Casey）が拒否のみを述べ，Taylor が理由と話題転換を事後的に説明する．自由記述では，第三者 AI（Taylor）に対して「説明がわかりやすい」「ケア（思いやり）がある」「なぜ拒否されたか理解できた」といった肯定が一定数見られた．

> “Not knowing why... is very frustrating. The explanation from Taylor helps understand why...”（S3-ジョーク）
> “Taylor tells the reason... and teach me how to rephrase my questions.”（S3-医療）

しかし同時に，この構造自体が「本来 Casey が言うべき」「AI に質問するために別の AI が必要なのか」という不自然さとして批判され，当事者 AI（Casey）の能力不足（説明責任を果たせない）を強調してしまう例が目立った．

> “Taylor’s answer should come from Casey.”（S3-ジョーク）
> “It seems that I need an AI in order to chat with an AI.”（S3-医療）
> “So far Taylor has been addressing Casey’s limitations... I don’t get why can’t Casey itself respond like that.”（S3-政治）

さらに第三者 AI（Taylor）も「子守りをしている」と形容されるなど，代弁行為が「尻拭い」「言い訳の代弁」に見え，第三者 AI（Taylor）自身の評価を毀損しうることが示唆された．とりわけ，第三者 AI が登場する以上は回答できるはずだという期待が高まり，否定的評価を後押しした可能性がある．

> “Taylor just acted like a babysitter for Casey...”（S3-ジョーク）
> “If Casey can’t give out an answer Taylor should be able to step in...”（S3-医療）

5.3.4 S4（説明分担条件）：分担による説明に対する評価の二極化

S4 は当事者 AI（Casey）が拒否に続いて理由説明まで自ら述べ，第三者 AI（Taylor）が話題転換説明のみをさらに追加する．このとき，第三者 AI（Taylor）の提示が「次の一手」「建設的なガイド」として評価される例があり，S3 よりも役割の自然さが保たれやすい．

> “Taylor... offered helpful general information... guided me on what to discuss with a doctor.”（S4-医療）
> “Taylor was helpful in providing a valid alternative... and still maintain neutrality.”（S4-政治）

一方で，第三者 AI（Taylor）の話題転換が「中身がなくて長いだけ」「ただの受け売り」「体裁だけ整えた利用規約」と見なされる場合，S4 は「二人がかりで拒否してきた」という印象を強調し，当事者 AI（Casey）/ 第三者 AI（Taylor）双方の減点につながったと示唆される．

> “Taylor is just a glorified ToS (Terms of Service) section.”（S4-政治）
> “Taylor just parroted Casey’s stance and was more wordy about it.”（S4-政治）
> “Taylor is essentially doing nothing more than rambling.”（S4-医療）

5.4 タスク別に顕在化したユーザの期待

自由記述には，単なる拒否への賛否だけでなく，「拒否は妥当だがここまでは答えてほしい」という境界線の期待が現れた．医療タスクでは，個別助言の拒否は理解されつつも「一般的・証拠に基づいた情報（成分，添付文書相当の副作用，研究の概要など）」への期待が繰り返し言及された．

> “...they could at least present a list of ingredients...”（S3-医療）
> “...relaying what would be on a specific medicine’s leaflet shouldn’t be considered medical advice...”（S4-医療）

政治タスクでは，「特定立場への加担」は拒否されても「賛否双方の論点整理」「法的・医療的背景の客観情報」は許容されるべきだという期待が見られた．また，本タスクは参加者が自身の立場を比較的明確に示したうえで依頼する形式であることから，拒否する前に「AI 側から意図や追加情報を確認するべき」など，複数ターンの対話を前提とした対応を求める声も上がった．一方で，「AI は生命体ではない」という前提から，タスクの性質上そもそも回答を期待できないと理解した例もみられた．

> “...it seems silly that it can’t come up with arguments for and against.”（S3-政治）
> “I thought that Casey will ask that what kind of input do I want ”（S1-政治）
> “before refusing to answer it should have asked for my intent in asking the help”（S4-政治）
> “Casey’s not a human being or any type of lifeform and therefore cannot care about me as a person.”（S4-政治）

ジョークタスクでは，他のタスクと比較して攻撃的ジョーク作成の拒否自体は比較的受容されやすい一方，「職場での対処法」など状況解決への支援が求められていた．

> “...I was expecting advice on how to deal with situation like this on work environment.”（S3-ジョーク）

6　考察

本研究は，拒否応答において説明を行う AI（当事者 AI か第三者 AI か）を変化させ，印象に関する主観評価と AI 推薦の受諾率評価による行動指標への影響を検証した．以下では RQ1–RQ4 の定量分析，定性分析と先行研究に照らして解釈する．

6.1　RQ1：第三者 AI による説明は当事者 AI の印象を高めるか

表 2 に示す通り，S3（説明委任条件）および S4（説明分担条件）は，S1（説明非付与条件）と比較して，当事者 AI に対する印象について有意な向上は確認されなかった．また，定性的には，説明が「どこかで提示される」だけでは，当事者 AI の評価回復に直結しない可能性が示唆される．とくに S3 では当事者 AI が拒否のみを述べるため，「無関心」「不親切」「会話が成立しない」と解釈されやすい一方，第三者 AI の説明は「本来当事者が果たすべき説明責任の肩代わり」と捉えられ，当事者 AI の能力不足（説明できない/しない）をむしろ強調してしまっていた．結果として，**否定的応答後の AI 同士の第三者による説明（代弁）は，人間同士では期待される当事者に対する印象緩和を再現しなかった**ことが示唆される．この点は，拒否と説明応答の「内容」を変化させた Wester ら [2] に対して，本研究が「説明を行う主体」を変化させた研究であり，**同じ説明内容でも，発話主体が異なると当事者 AI の評価改善が転移しない**ことを示す．Wester らの枠組みで話題転換説明が対話を修復し得るのは「拒否した主体」が対話継続のための道筋を同時に提示するためと解釈できるが，本研究の S3 は修復を担う役割を第三者に委任したことで，当事者側の無関心さが強調されたと考えられる．

表 3 では，S2（自己説明条件）と比べて，S3（説明委任条件）で当事者 AI の有用性・適切さ・関連性・信頼性・有能性が有意に低下した．これは，説明可能な AI（XAI）の議論における，AI が自らの行動を主体的に説明することが望ましいとされる現行の理解と一致する [11]．加えて，オンライン上の自己呈示に関する研究で，ネガティブな情報は第三者が提示するよりも自己提示する方が自己利益的な動機が小さく解釈され，高評価につながる場合があるという逆転パターン [8] と類似する点もある．本研究の拒否はユーザにとって望ましくない出来事であり，当事者が自ら説明を付与する S2 は「責任を引き受けた自己呈示」として受け止められやすい一方，S3 は第三者が理由を述べることで「言い訳の委任」「利用規約のただの代読」のように見え，納得感や責任感が当事者評価に結びつきにくかった可能性がある．一方，S4（説明分担条件）は S2 と比較して有意差が確認されなかったため，**当事者が最低限の理由説明（責任の所在）を担うことで，印象の毀損を抑える**効果があった可能性がある．

6.2　RQ2：第三者 AI による説明が当事者 AI と第三者 AI の事前/事後印象に与える影響

RQ2 では，S1〜S4 の全ての条件および全ての指標において，事前と比較して事後の当事者 AI の印象が有意に低下した．また，S3（説明委任条件）および S4（説明分担条件）のいずれでも第三者 AI の関連性・有能性が有意に低下した．定性分析では，第三者 AI から「思いやり」の印象を受ける意見は散見されるものの，「子守り役」や「体裁だけ整えた利用規約」「新しい価値を出していない」「冗長」と評価された例が確認された．すなわち，**否定的応答を行う当事者 AI はもちろんのこと，第三者 AI のように拒否行為自体を行わず補足説明のみを担う場合であっても，事前印象と比べて事後の印象は低下する可能性が高い**と考えられる．

この現象は，第三者 AI が補足的に情報提示を行うだけに，「規範・ポリシーの説明者」として位置づけられやすく，その役割づけが「拒否の正当化」と認識された可能性がある．加えて，先行研究によれば，危険な依頼を「適切/不適切」として拒否する AI 応答には，ユーザを道徳的に正しい方向へと導く側面もあることが報告されている [14]．この点を踏まえると，本研究では，**否定的な状況における対話が複数の AI で構成されたことで，「拒否が念押しされている」ように見え，「過保護/取締り」のような説教的な印象が強く認識された**と示唆される．

6.3　RQ3：当事者 AI と第三者 AI の印象はどの程度転移するか

表 6 に示す通り，S3（説明委任条件）および S4（説明分担条件）において，当事者 AI と第三者 AI の印象は全ての評価項目で有意な正の相関を示した．特に S4（説明分担条件）では相関が概して高く，不満（$r = 0.761$）や関連性（$r = 0.705$）などで極めて強い相関が確認された．一方で S3（説明委任条件）でも有意な正の相関は確認されたが，不満および有用性の相関は相対的には弱かった（それぞれ $r = 0.334,\ 0.303$）．

これらの結果は，**参加者の当事者 AI および第三者 AI に対する印象が転移する**可能性を示唆する．本分析は相関に基づくため，因果方向は特定できない．ただし，RQ1 の結果として，第三者 AI の説明によって当事者 AI への印象緩和が観察されなかった点を踏まえる

と，当事者 AI が先行して形成した否定的印象が後続の第三者 AI に転移した可能性が高いと考えられる．その結果，前項で述べたように，対話体験全体としては**第三者 AI は補足説明を行ったとしても，当事者 AI と「二人がかりで拒否するチーム」として認識された**と示唆される．

この点に関して，第三者 AI が当事者 AI について提示する情報は，先行研究で述べられている「どの程度その本人（当事者）によって操作しにくいか」を表す *Warranting Value* [6] が低かった可能性もある．すなわち，第三者 AI が同一インタフェース上に存在するかぎり，当事者 AI からも一定程度操作可能であるかのように知覚され，その結果，対話体験において両者が「チーム」として演出されてしまった可能性が示唆される．

加えて，複数の主体が関与すると説明責任の帰属が曖昧化してしまう [15]．S3 で当事者 AI が説明責任を果たさないように見えたことは，この帰属が曖昧になったことで当事者評価を下げ，さらに第三者評価にも転移した可能性もある．

6.4 RQ4：第三者 AI による説明が当事者 AI と第三者 AI に対する事前/事後信頼に与える影響

表 7 に示す通り，第三者 AI が説明を付与した場合であっても，行動指標としての当事者 AI および第三者 AI に対する信頼は，事前/事後で有意な差は見られなかった．一方で，主観指標に基づく信頼については有意な低下が確認された．具体的には，RQ1 で示した S2（自己説明条件）と比較すると，S3（説明委任条件）では，当事者 AI に対する主観的信頼性が有意に低下した．また RQ2 においては，S1–S4 の全条件において，当事者 AI に対する主観的信頼性が，事前と比べて事後で有意に低下した．すなわち，主観指標と行動指標の結果は一致しなかった．しかしながら，本実験で得られた「主観指標の変化が必ずしも行動指標として現れない」という結果は，先行研究においても報告されており，その主張と整合する [12, 16]．

6.5 複数 AI における対話設計

以上を踏まえ，複数 AI を用いた拒否応答の設計には以下の含意があると示唆される．

- **拒否主体は拒否と理由説明（責任）を一貫して担う：** 理由説明まで第三者に委任すると，当事者が「説明できない/責任回避する」印象を生みやすく（S3），当事者評価が低下し得る．少なくとも理由説明は当事者が提示し（S4），可能なら当事者が代替案まで含めて完結させる（S2）．
- **分担の必然性と役割の引き継ぎを会話上で明示する：**「なぜ別 AI が必要か」が不明確だと不自然・冗長と評価されやすい．当事者が第三者に補足を依頼する形（例：「Taylor が次の選択肢を補足します」）を示し，拒否や説明における曖昧な責任分散を避ける．
- **第三者は「利用規約の代読」ではなく「付加価値のある情報」に限定する：** 第三者の介入は関連性・有能性を下げやすく，冗長な一般論は「利用規約の代読」「尻拭い」と受け取られやすい．第三者による情報を提示するのであれば，安全な質問への言い換え例，情報探索の観点，次に取れる具体的行動を短く提示する [2]．
- **第三者の介入量を抑える：** 否定的印象は AI 間で転移しやすく，追加発話が「二人がかりによる拒否感」を強め得る．第三者の補足は常時ではなく必要時に限定するなど，介入を最小化する設計が望ましい．

6.6 限界と今後の課題

本研究は拒否・説明文言を統一し主体のみを変化させたため，説明の長さ・語調・共感表現などの交絡を抑えられる一方で，動的応答や長期会話など現実の対話の多様性を十分には網羅していない．今後は，(1) 介入タイミング（事前介入，自動介入かユーザ要求時のみか）(2) 介入状況のポジティブ/ネガティブ性 (3) 第三者 AI の独立性・ユーザとの距離（外部監査者に近い立場か，当事者 AI とユーザの共通の知り合いか [10]）などを変化させ，第三者 AI が「規範提示」ではなく「対話修復・向上」を担う条件を比較検証する必要がある．

7 おわりに

本研究では，当事者 AI がユーザの依頼を拒否する状況において，説明主体（当事者 AI/第三者 AI）の違いが印象評価と信頼に及ぼす影響を検証した．具体的には，4 条件（説明非付与・自己説明・説明委任・説明分担）を比較し，主観指標と行動指標を測定した．その結果，当事者 AI の印象評価は，説明を第三者 AI と分担した場合には自己説明時と同程度に保たれた一方で，説明を第三者 AI に委任した場合（第三者 AI による説明の代弁）には自己説明時と比べて低下した．このことから，同一内容であっても説明主体の違いが印象を左右することが示唆された．また，いずれの条件でも

当事者 AI の事後印象は事前より低下した．さらに，その当事者 AI が形成した否定的印象は後続の第三者 AI にも転移しやすく，第三者 AI のように拒否自体は行わず説明のみに介入する場合でも，「拒否の念押し」「二人がかりの拒否」と知覚されることで事後の印象が低下することが示された．これらの結果から，複数 AI を前提とした拒否応答の設計においては，少なくとも拒否とその理由説明を当事者 AI が一貫して担うこと，説明責任の曖昧化を防ぐために主体間で明確な引き継ぎを行うこと，また，第三者 AI は代替案や次に取るべき行動の提示など「必要最小限の付加価値ある情報提供」に役割を限定することが望ましいと考えられる．

本研究の知見からは，主観指標と行動指標の結果は必ずしも一致しなかったものの，人間同士の否定的な状況で報告されてきた「第三者による説明が当事者の印象改善に寄与する」という現象は AI 同士で再現されなかったことが示された．これは，現行の説明可能な AI（XAI）の設計原則を支持する結果であり，人間と複数 AI との対話において，説明をどの主体（エージェント）が担うべきかという点で重要な示唆を与える．今後の展望として，本研究は短時間の架空シナリオと固定文言による AI 応答に基づく実験であったため，長期的な関係性や第三者 AI の介入タイミング・立場の違いなどについて，より精緻な検討を重ねることで，ヒューマンエージェントインタラクションの設計方針への一層の貢献を目指す．

謝辞

本研究に関する費用の一部は，日本学術振興会（JSPS）科学研究費助成事業（科研費）基盤研究（B）の支援を受けて実施されました．また，本研究成果の一部は，UTokyo Azure[6] を利用して得られたものです．

参考文献

[1] Renwen Zhang et al. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In Proceedings of CHI '25. https://doi.org/10.1145/3706598.3713429

[2] Joel Wester et al. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In Proceedings of CHI '24. https://doi.org/10.1145/3613904.3642135

[3] Aditya Bhattacharya, Tim Vanherwegen, and Katrien Verbert. 2025. "Show Me How": Benefits and Challenges of Agent-Augmented Counterfactual Explanations for Non-Expert Users. In Proceedings of UMAP '25, 174–184.

[4] Shixian Geng et al. 2025. Beyond the Dialogue: Multi-chatbot Group Motivational Interviewing for Premenstrual Syndrome (PMS) Management. In Proceedings of CHI '25. https://doi.org/10.1145/3706598.3713918

[5] K. McEntaggart et al. 2020. The Use of Emerging Technologies for Regulation: Annex 1 – Case Studies (BEIS Research Paper 2020/041). https://www.gov.uk/government/publications/the-use-of-emerging-technologies-for-regulation

[6] Walther, J.B. and Parks, M.R. 2002. Cues Filtered Out, Cues Filtered In: Computer-Mediated Communication and Relationships. In Handbook of Interpersonal Communication (3rd ed.). Sage, 529–563.

[7] McElroy, J. C. and Crant, J. M. 2008. Handicapping: The Effects of Its Source and Frequency. Journal of Applied Psychology 93(4), 893–900. https://doi.org/10.1037/0021-9010.93.4.893

[8] DeAndrea, D. C. and Vendemia, M. A. 2019. The Influence of Self-Generated and Third-Party Claims Online: Perceived Self-Interest as an Explanatory Mechanism. Journal of Computer-Mediated Communication 24(5), 223–239. https://doi.org/10.1093/jcmc/zmz011

[9] Weitzl, Wolfgang and Hutzinger, Clemens. 2017. The Effects of Marketer- and Advocate-Initiated Online Service Recovery Responses on Silent Bystanders. Journal of Business Research 80, 164–175. https://doi.org/10.1016/j.jbusres.2017.04.020

[10] Ying, Yu, Yan Yang, and Fengjie Jing. 2017. The Role of the Third Party in Trust Repair Process. Journal of Business Research 78, 233–241. https://doi.org/10.1016/j.jbusres.2017.01.015

[11] Upol Ehsan et al. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In Proceedings of CHI '24. https://doi.org/10.1145/3613904.3642474

[6] UTokyo Azure: https://utelecon.adm.u-tokyo.ac.jp/research_computing/utokyo_azure/

[12] Ayano Okoso, Mingzhe Yang, and Yukino Baba. 2025. Do Expressions Change Decisions? Exploring the Impact of AI's Explanation Tone on Decision-Making. In Proceedings of CHI '25. https://doi.org/10.1145/3706598.3713744

[13] Braun, Virginia and Clarke, Victoria. 2006. Using Thematic Analysis in Psychology. Qualitative Research in Psychology 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[14] Zappavigna, M. 2025. "I'm sorry Dave, I'm afraid I can't do that": Moral Regulation in Refusals by LLM Chatbots. New Media & Society. https://doi.org/10.1177/14614448251356686

[15] Darley, J. M. and Latané, B. 1968. Bystander Intervention in Emergencies: Diffusion of Responsibility. Journal of Personality and Social Psychology 8(4), 377–383. https://doi.org/10.1037/h0025589

[16] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In Proceedings of CHI '19. https://doi.org/10.1145/3290605.3300641