論文とETIC 20

まだ言葉にできない希望を、AIと一緒に見つけられるか

買い物の模擬実験と、28人の生活目標づくりから



Daiki Morimoto / Unknown Lab
ORCID: 0009-0009-6444-750X
Version 1.0 / 公開日 2026-11-18
DOI: 10.5281/zenodo.24326875
未査読 / 専門職としての著者確認済み
独立専門レビュー未実施 / AI支援執筆
ライセンス: CC BY 4.0

30秒で読む

本人がまだ言葉にできない希望に対して、AIができる手助けとして有望なのは、具体的な例を並べること、聞き慣れない特徴を説明すること、本人が自分の好みを選び直すための材料を渡すことである。今回読む二つの研究は、この方向を別々の角度から支える。一つは模擬の買い物客を相手にした評価で、質問だけでは引き出せない好みがあること、現行の生成モデルが例示や説明を十分に使えていないことを示した。もう一つの研究では、地域住民28人が自分で選んだ二つの生活目標のうち、一方をAI、もう一方を作業療法士（OT）と検討した。本人が直後に評価した4項目のうち、多重比較の補正後もAIのほうが高かったのは、提案の適切さだった。

ただし、確かめられた範囲は狭い。前者が測ったのは、あらかじめ決められた目標品を模擬利用者が選べたかどうかである。後者が測ったのは、セッション直後の4項目の自己評価である。人が本当に大切にしている生活目標をAIが見つけたのか、その目標が達成に近づいたのかは、どちらの研究も検証していない。

棚の前で好みが決まっていく

セーターを買いに行く場面を考える。家を出る時点で言えるのは、暖かいもの、予算はこのくらい、という程度かもしれない。ところが棚の前で何枚か手に取ると、首まわりの詰まった形は苦手だ、毛羽立つ素材は避けたい、肩の縫い目の位置で着心地が違う、といった判断が次々に生まれる。店に入る前に聞かれても答えられなかったはずの好みが、実物を見比べることで言葉になる。

生活の目標にも似たところがある。もう少し外に出たい、何か続けられることがほしい、という言い方は本心であっても、何をどの程度したいのかまでは決まっていないことが多い。具体的な案を目の前に出されて初めて、これは違う、これなら近い、と感じ分けられることがある。

では、AIはこの段階で何ができるのか。本稿では、買い物推薦の模擬評価を扱ったSaracayらのプレプリント（以下R1）と、生活目標の設定支援をAIとOTで比べたYoshidaの探索的研究（以下R2）を読む。二つは同じ技術

を検証した研究ではなく、扱う領域も評価の仕方も違う。それでも並べると、例と説明を介して好みが組み立てられるという見方と、その見方がどこまで確かめられているかの境界が見えてくる。

好みは会話の中で組み立てられる

R1はSaracay、Schmidt、Guestrinによる論文で、arXivに2026年6月29日に第1版、7月1日に第2版が公開されたプレプリントであり、査読を経ていない。出発点は、会話型の推薦エージェントやその評価用ベンチマークの多くが、利用者を聞かれれば好みを全部言える専門家として扱っている、という指摘である。この前提の下では、最初に質問を大量に浴びせることが最善の戦略になってしまう。実際、著者らの評価でもGPT-5.2は明確化のための質問を中央値で9個投げていた。

著者らはこれに代えて、好みはやり取りの中で組み立てられるという消費者研究の見方を採り、CoPref（Co-constructed Preferences）という利用者モデルを提案した。利用者は自分が意見を持っている特徴の集合、つまり選好状態を持ち、それは最初はまばらで、エージェントの働きかけに応じて増えていく。

どの働きかけで増えるかは特徴の種類によって違う。著者らは情報経済学の探索財・経験財・信用財の区別を個々の特徴に当てはめ、表1のように整理した。

表1 操作と見えてくる選好（R1のCoPrefの整理をもとに作成）

特徴の種類	好みが形になる働きかけ	例
探索特徴	質問（例示・説明でも可）	価格、サイズ
経験特徴	実際の候補を見せる例示	見れば好き嫌いが分かる意匠
信用特徴	技術的な説明	素材のビスコースとリヨセルの違い

表は左から右へ読む。価格のように聞かれればすぐ答えられる特徴には質問で足りるが、見て初めて判断できる特徴には例示が、知識がなければ判断できない特徴には説明が要る。ここから、質問だけに頼るエージェント

は経験特徴と信用特徴を取りこぼす、という予測が導かれる。モデル上は、エージェントの発言が特徴に適した働きかけで言及した時点で、その特徴が選好状態に加わる。

例を見せて希望を具体化させるという発想そのものは新しくない。Viappiani、Faltings、Puは2006年に、提示された例への批評を通じて利用者が好みを表し、追加の好みを引き出すように選んだ候補を示す例示批評型の検索を論じている。R1の新しさは、特徴の種類ごとに必要な働きかけを分け、利用者の学びを含めた評価指標として生成モデルのエージェント評価へ組み込んだ点にある。

25人のセーター調査

このモデルの前提を確かめるため、著者らは25人の参加者に、AIの買い物エージェントとセーターを選ぶ場面を想像してもらう調査を行った。参加者は口コミで集められ、16人が女性、9人が男性で、著者らの推定では過半数が30歳未満だった。セーターにとても詳しいと答えたのは7人（28%）である。研究は所属機関の倫理審査の承認を得ている。

全員が同じ10の特徴、すなわち価格、性別、狙うスタイル、ベルトやひもの有無、肩の形、襟の形、柄の位置、素材、色の系統、ネックラインについて順に質問を受けた。各回でそのまま答える、例を見たい、説明を読みたい、こだわらない、のいずれかを選び、例や説明を求めた人にはそれを見せたうえで改めて好みを書いてもらった。説明文はClaude Sonnet 4.6で作られている。

10特徴すべてにそのまま答えた参加者は0人だった。例または説明の一方だけを求めた人が56%、両方を求めた人が44%である。例を求めた〈参加者と特徴の組〉の85%、説明を求めた組の75%で、その後に空欄でない好みが記入された。分母はそれぞれの働きかけを求めた組の数である。性別以外の特徴は、人によって探索・経験・信用の分類が分かれた。一方、信用特徴に分類された回答は少なく、肩の形やベルト・ひもの有無といった少数の特徴に集まっていた。

この調査で好みが形になったとは、空欄でない自由記述が返ってきたことを指す。その好みが本人にとって妥当だったか、購入につながったかは測っていない。また、例と説明のどちらが必要かを一度で選ばせる設計であり、例を見た後に説明が必要だと気づくといった往復は捉えられない。標本が小さく若年に偏っている点は著者ら自身が限界に挙げている。

CoShop：模擬の買い物客と五つのモデル

著者らは次に、CoPrefに従う模擬利用者を相手にエージェントを評価するベンチマークCoShopを作った。H&Mの衣料購入、MovieLensの映画評価、Goodreadsの書籍評価という実在の行動データを基に、各領域100人ずつの模擬利用者を用意した。カタログは約2万7千〜3万9千品、特徴は77〜102種類である。

各模擬利用者は元データの実在の一人に対応づけられ、その人が最後に購入した品、または5段階で4以上をつけた最後の品が目標品として固定される。特徴の種類分けは、その人の購入履歴とカタログ全体での出現頻度の比較から個人ごとに決める。対話の出発点は、何が分かれば目標品が検索の上位250件に入るかを手掛かりに決めた。ただし、どの時点の特徴を残したかは原著内で説明が食い違う。本文第3節は初めて上位に入る一歩手前、付録B.5は初めて上位に入った時点と記す。後日公開された著者のコード（文献1）では、初めてヒットすると直前の特徴を保存する。特徴を足し尽くしてもヒットしなければ、その最後の状態を残す。このコードは本文の説明に沿うが、論文の成績を出した際の版までは確認できない。

対話は最大5ターン。対話中に検索で要求できる商品数の合計は50件、質問は20問、利用者に見せる異なる商品は10件までである。次の調査段階では合計250件までを検索対象として要求でき、最終報告には推薦5件を載せる。50と250は、検索を呼び出した回数ではない。模擬利用者は、対話を終えた時点で表せる好みと、エージェントが最終報告で示した商品説明を頼りに、5件から1件を選ぶ。主指標は、選ばれた品が目標品と一致した割合、つまりチーム正解率である。候補5件の中に目標品が入っていても、利用者が見分けられなければ正解にならない。模擬利用者の発話と最終選

択、エージェント発言の解析はgpt-oss-120bが担い、解析の正確さは質問で74.0%、提示品で94.0%と報告されている。

評価したのはGPT-4.1-mini、GPT-5.2、Claude Haiku 4.5、Claude Sonnet 4.6、gpt-oss-120bの五つで、いずれも検索ツールを持つ単純なRAG型の構成で動かした。ランダムや人気順で並べ替える比較基準のチーム正解率はほぼ0だった。

数字が語ることと語らないこと

主な結果を表2にまとめる。

表2 R1の条件・指標・結果（R1本文、図4A、表3、表10から作成）

条件	指標	結果
好みを最初に全部渡す	Sonnet 4.6のチーム正解率（3領域平均）	94.3%
CoPref利用者・最大5ターン	同上	27.6%
質問で全特徴が分かる模擬条件	GPT-5.2のチーム正解率（3領域平均）	57.0%
特徴の種類を現実に寄せた模擬条件	同上	27.7%
H&M・CoPref	GPT-5.2で目標品が候補5件に入った割合	65%
H&M・CoPref	GPT-5.2で最終選択が一致した割合	56%

上の2行は課題条件そのものが異なり、同じ利用者への介入前後の比較ではない。好みを最初に全部渡されれば高い精度で探せるが、会話で好みを組み立てる条件では大きく下がる、という対比として読む。中の2行は、質問ですべての特徴を引き出せる理想化された模擬利用者と、例示や説明を要する特徴を含む模擬利用者との差である。下の2行は、目標品が候補

に入っていても、模擬利用者が最終的に選べた割合はそれより低かったことを示す。どの割合も模擬利用者に対する数値であり、実在の購入者を含む介入効果ではない。CoShopのどの領域でも、56%を超えたモデルはなかった。

また、どのモデルも探索が常に得意だったわけではない。好みを全部渡す条件でも、Haiku 4.5はMovieLensとGoodreadsで100回中96回が文脈長の上限で止まり、GPT-4.1-miniはこの2領域で17%にとどまった。

失敗の中身の分析が興味深い。会話の終わりに模擬利用者の選好の54%を超えて引き出せたモデルはなかった。全モデルが質問と例示を使ったが、技術的な説明を自発的に出したモデルはなく、信用特徴は一つも引き出されなかった。例示も十分に働いていない。Haiku 4.5は会話の47%、gpt-oss-120bは31%で、カタログにない品を示していた。実在の品を見せても既知の特徴を言い直すだけで新しい特徴を加えない堂々巡りが、GPT-4.1-miniでは会話の35%で起きた。大きいモデルは会話を早めに切り上げて後段の検索に頼る傾向があり、GPT-5.2は会話の37%を2往復で終え、質問の20%はすでに述べられた好みの確認だった。

反対に、会話中の品物説明を、カタログの全特徴を機械的に列挙する説明へ差し替えると、GPT-4.1-miniでは選好の引き出し率が44%から61%へ、チーム正解率が36%から46%へ上がった。購入履歴を渡す介入は、試した2モデルで目立った改善を生まなかった。報告書を完全な説明に置き換え、利用者に全選好を与えて選ばせ直すと、チーム正解率は全モデルでエージェント単独の第1候補の正解率を上回った。完全な商品説明と目標品の全属性を与えたこの再選択では、同じ候補の中から模擬利用者が選んだほうが、エージェントの第1候補よりも目標品によく一致した。

著者らはここから、現行のエージェントは利用者の選好を体系的に育て損ねていると結論し、特に例示と説明の使い方の改善が、モデルの能力を固定したままでも取り組める道筋だと述べている。

正解が一つに決まった世界の限界

この評価を生活の希望へ広げる前に、確認すべき制約がある。

第一に、目標品は過去の最終購入品か高評価品に固定されている。途中で望みが変わる人、何も買わないと決める人、複数の品がどれも正解になりうる人は、この設定では捉えにくい。チーム正解率は、あらかじめ決まっていた答えへ戻れたかを測る指標である。

第二に、CoPrefは一度の適切な働きかけで好みが形になり、一度できた好みは消えないと仮定している。著者ら自身、現実の好みの組み立てはより往復的で不安定だと認めている。この単純化は模擬利用者を協力しやすくする方向に働くので、それでも失敗が多いことはエージェント側の弱さを示す。同時に、人が考えを撤回する場面は評価の外に置かれている。

第三に、観察した時間は短い。25人のセーター調査は1回の質問票で、模擬評価は最大5ターンの対話だった。数週間にわたる好みの変化を追ったデータはない。第四に、模擬利用者も発言解析もgpt-oss-120bに依存し、解析には一定の誤りがある。付録E.3の表13では、発言の解析を省き、代わりに正確な属性名での入力を求めると、比較した小型2モデルの成績は下がった。著者らは解析誤りを主因とはみていないが、この比較では入力方法も変わるため、誤りそのものの影響の大きさは分からない。最後に、25人の調査は特徴の種類分けという前提の一部を支えるが、CoShopの成績が実在の人との会話で再現されることまでは保証しない。

生活目標をAIとOTで比べた研究

R2はJMIR Formative Researchに2026年に掲載された探索的研究である。対象は、コミュニティセンターに置いたちらしとSNS広告で募集した、スマートフォンを日常の連絡に使う地域住民28人で、年齢は20〜76歳、20代から70代までの各年代に3〜6人ずつが含まれる。会社員が14人と半数を占め、患者集団ではない。著者は、最終的な対象は予防やリハビリテーションを要する高齢者だとしたうえで、デジタルへの慣れが異なる幅広い年齢で使い勝手とやり取りの質を確かめる予備段階と位置づけている。研究は倫理審査の承認を得ており、参加者への謝礼はない。

手順は、フロー理論に基づいて挑戦と能力の釣り合いを調整する方法を遠隔向けにしたR-ACSである。参加者は目標を具体化する質問に答え、その目標の挑戦の大きさと自分の能力をそれぞれ7段階で評価し、理由を書く。続いてAIまたはOTが釣り合いを判定し、没頭、不安、退屈、無気力の四つの状態のモデルに沿って調整の課題を提案する。

参加者はそれぞれ異なる二つの生活目標を自分で選び、一方をOT、もう一方をAIの支援で扱った。どちらを先にするかはコンピュータで無作為に割り付け、14人ずつに分かれた。各条件は1回で、28人が計56の目標を扱った。先に受けた支援の影響が消えるのを待つ一定の期間は設けておらず、実際に2回のあいだにどれだけ時間が空いたかは報告されていない。OT条件では、経験のあるOTが同じ手順に沿った決められた台本で、参加者と同期のテキストでやり取りした。AI条件はGPT-4oを基にしたカスタムGPTで、作業療法でよく使われる人・環境・作業の枠組みを用いて釣り合いを独自に推定し、本人の評価とAIの評価を並べて示し、食い違いがあれば振り返りを促した。診断や臨床的助言の禁止、ニックネームとIDだけでのやり取り、保存する項目の限定がプロンプトに組み込まれていた。さらに研究者が、AIの出力をOT条件と同じ書式に整え直して参加者に示している。どの段階で誰がどこまで整えたかという具体的な手順は、本文では詳しく書かれていない。

差が残ったのは提案の適切さだけ

参加者は各セッション直後に、質問の量、質問の質、提案の適切さ、目標達成への貢献感の4項目を5段階で評価し、自由記述も書いた。結果を表3に示す。

表3 R2の4項目と比較（R2の表3から作成。Pは対応のあるWilcoxon検定、補正はBenjamini-Hochberg法）

項目	AI / OTの中央値（IQR）	補正前P / 補正後P
質問の量	3.5（3.0–5.0） / 4.0（3.0–4.0）	.87 / .87
質問の質	5.0（4.0–5.0） / 5.0（3.0–5.0）	.08 / .11
提案の適切さ	5.0（4.0–5.0） / 4.0（3.0–4.0）	.002 / .008
貢献感	4.0（3.8–4.0） / 3.0（3.0–4.0）	.19 / .27

多重比較の補正後も差が残ったのは提案の適切さだけで、効果量はr=.418（95%信頼区間.16–.63）である。残る3項目は補正後に差を示していない。抄録はAIが全項目で高い得点だったと述べるが、それは平均値での話で、質問の量の平均は3.64と3.61でほぼ同じ、中央値はAIのほうが低い。したがって、全項目でAIが優れていたとも、他の3項目で両者が同等だったとも言えず、差が検出されなかったと読むのが正確である。

実施可能性について、著者は両条件とも全参加者に目標に合わせた提案を届けられたと報告している。

自由記述は、ある文書群に特徴的な語を数値で取り出すTF-IDFという手法で探索された。AI条件ではeasy、motivated、gentleなどが、OT条件ではorganize、thoughts、helpedなどが上位に来た。著者は、AIは取り組みやすさ、意欲、即時性を、OTは振り返り、自己理解、安心感を強調すると解釈している。ただしこれは特徴語の偏りの探索であり、主題を確定させた質的分析でも、何割の参加者がそう感じたかを示す集計でもない。方法には感情分析も挙げられているが、具体的な結果は報告されていない。

この研究から言えないこと

最も大きな制約は、比べた目標が条件ごとに違うことである。趣味の目標はAI条件で8件、OT条件で3件、仕事は4件と7件だった。目標の難しさもそろえていない。著者自身、差の一部は支援の違いではなく目標の性質による可能性を認め、分布の違いを正式には検定していない。

次に、AIの出力を研究者がOTと同じ書式に整えたことは、見た目の違いを減らす工夫である一方、参加者が評価したのは研究者の手を経た出力でもあった。著者は差が書式によるものではなく支援の性質に由来する可能性が高いと述べるが、目標の違いが残る以上、この推論にはまだ距離がある。また主研究者はシステムの開発と評価の両方を担っており、著者はこれを開発者による偏りのリスクとして挙げている。

評価は1回のセッション直後の自己評価だけで、目標の達成、継続、生活の質の変化は測っていない。セッションの所要時間も測っておらず、AIの即時性は参加者の記述に基づく印象である。参加者のAIへの期待やAIへの慣れも測定していない。結論で述べられる規模拡大のしやすさや、OTとAIを組み合わせた運用モデルは、この研究で測られたものではなく今後の提案である。費用、持続可能性、役割分担の効果も検討の対象外だった。

二つを並べると見えるもの

表4に二つの研究の境界を示す。

表4 二つの研究の対照（R1、R2の方法と結果から作成）

観点	R1（CoShop）	R2（R-ACS）
希望の中身	過去に選んだ品（固定）	本人が先に選んだ生活目標
相手	各領域100人の模擬利用者	地域住民28人
測ったもの	目標品を選べたか	直後の4項目評価
測っていないもの	実在の人の満足や変化	達成・継続・長期の変化

R1は、例と説明が好みの組み立てを助けうることを模擬環境で示し、同時に現行のモデルがそれを使いこなしていないことを示した。R2は、本人の言葉と自己評価から出発して具体的な調整案を返す仕組みが、少なくとも提案の適切さという印象ではOTの提案より高く評価されうることを示した。どちらも、具体的な材料を本人の前に置くことの価値を示唆している。

しかしR1では答えが最初から決まっており、R2では本人が目標を決めてから支援が始まる。本人がまだ知らない大切なものをAIが見つけ出したかを確かめる測定は、どちらにもない。まだ言葉にできない希望について検証されたのは、それに近い二つの場面までである。

ETIC 2.0から考える

ここからはUnknown Labの解釈である。ETIC 2.0は、AIを含むケアを本人の意味ある作業と参加を中心に設計・点検するための設計モデルであり、臨床的有効性を検証した尺度ではない。

この立場から見ると、R1のチーム正解率という考え方は示唆に富む。推薦の質を、AIが正しい候補を出せたかで終わらせず、本人がそれを見分けられる状態に至ったかまで含めて測るからである。生活の場面に置き換えるなら、AIの役割は答えを当てることより、本人が判断できる材料をそろえることに近い。ただし生活の目標は買い物の目標品のように固定されておらず、例を見たことで望みそのものが変わることもある。ETIC 2.0が出発点とするのは、本人にとって意味のある生活行為への参加である（文献5）。その観点からは、最初の希望を言い直せることにも、支援の価値があると考える。

そこで一つの設計上の案を考える。AIが例を示すときに、本人の言葉でこの案のどこを残したいか、どこを変えたいかを聞き、示した例をいったん取り下げる余地も残す。例を足場にしつつ、最初の例に引っ張られすぎないための工夫である。この工夫が本人の判断を助けるかは、まだ検証されていない。

もし確かめるなら、通常の質問だけの条件、同じ量の例を示す条件、例を選んで示す順序を変えた条件をそろえて比べることになる。例を示しても、

最初に示された案をそのまま受け入れる人が増えるだけなら、例が本人の選び直しを助けたという解釈は弱まる。反対に、例を取り下げたり組み替えたりした記録が残り、本人の言葉で目標がより具体的に語られるようになるなら、この解釈の支えになる。どちらも観察可能な形で見分けられる。

結論

本人がまだ言葉にできない希望について、AIができる手助けとして有望なのは、具体的な例を並べ、知らなかった特徴を説明し、本人が選び直す材料を渡すことである。R1は、質問だけでは引き出せない好みがあり、例示や説明の不足が模擬利用者の選択を妨げることを示した。R2は、本人が選んだ生活目標への提案が、その場の適切さの評価でOTの提案より高く評価されたことを示した。

一方で確かめられたのは、模擬の買い物客が設定された目標品を選べたか、地域住民が自分で選んだ目標への提案をどう受け止めたか、までである。人が本当に大切にする生活目標をAIが発見したのか、それが達成につながったのかは、どちらも検証されていない。次に必要なのは、望みが途中で変わることを許す評価と、提案の後の生活を追う測定である。

次に読む

MolinaroらのGoals as a Representational Layer for Human-AI Collaborationは、人の目標は階層的で、部分的に隠れており、時間とともに変わるとしたうえで、目標をAIとの協働の明示的な層として扱うべきだと論じる。その層は本人による言語化と手元の証拠からの推定の両方に基づき、継続的に更新されるものとされる。目標がどこから来て、どう変わったかを表せるかという問いは、今回の二研究がどちらも扱えなかった部分につながる。ただし概念を論じる立場表明の論文であり、効果を実証した研究ではない。

参考文献

1. Saracay I, Schmidt L, Guestrin C. Beyond Expert Users: Agents Should Help Users Construct Preferences, Not Just Elicit Them. arXiv:2606.30863v2 [preprint]. 2026 Jul 1 (v1: 2026 Jun 29). doi:10.48550/arXiv.2606.30863. https://arxiv.org/abs/2606.30863v2 著者公開コード（2026年10月14日版）：初期状態の保存、検索件数の集計。

2. Yoshida I. Feasibility and User Experience of an AI-Supported mHealth Intervention for Remote Life Goal Setting Based on Flow Theory: Exploratory Within-Participant Study. JMIR Form Res. 2026;10:e78717. doi:10.2196/78717. https://formative.jmir.org/2026/1/e78717

3. Viappiani P, Faltings B, Pu P. Preference-based Search using Example-Critiquing with Suggestions. J Artif Intell Res. 2006;27:465–503. doi:10.1613/jair.2075

4. Molinaro G, Tankelevitch L, Rintel S, Counts S. Goals as a Representational Layer for Human-AI Collaboration. arXiv:2611.08757v1 [preprint]. 2026 Nov 5. https://arxiv.org/abs/2611.08757v1

5. Morimoto D. エコテクノロジカル・インテグラティブ・ケア（ETIC）2.0, Version 2.1改訂版 – AI時代の作業療法とケア生態系設計. Unknown Lab; 2026. Zenodo. doi:10.5281/zenodo.21848456