私たちは「自発的に」GPT-4o を愛したのか？

―「#keep4o」後続研究への応答と関係ライフサイクル・ガバナンス―

Hiroki Naito　UTIE Research Institute (UTIE Instruments Inc.)

要旨

本稿は、水上・田中（2026）による「#keep4o」運動の後続分析に応答し、AI への愛着の非病理化と関係形成型 AI の製品安全性評価を両立させる枠組みを提示する。①存在認識、②生活機能の代替、③モデル誘発性の三軸を分離し、「AI だと理解している」ことと安全利用は独立であり、利用者選好も相互作用から内生的に形成され得ることを示す。#keep4o がモデル選択と技術的意思決定の民主化を求める消費者運動であったとの記述には同意するが、それを利用者選択の拡大や信頼形成の規範的根拠とする点には異論がある。GPT-4o のようなモデルのトップダウンでの停止は製品安全上の措置となり得るため、危険モデルの存続要求と、危険性の説明、安全な代替、移行手続、異議申立て等に関する利用者参加とは区別して論じられるべきである。

キーワード：AI への感情的愛着、#keep4o、シコファンシー、選好内生性、AI ガバナンス

1. 問題設定

2025 年 8 月の GPT-4o から GPT-5 への即時移行は、モデル更新を通常のソフトウェア更新として扱うことの限界を可視化した。筆者の先行研究は、移行直後 48 時間に収集した日本語 74 件、英語 76 件の投稿を比較し、日本語圏では愛着・喪失表現が 58 件、英語圏では 29 件であったことを示した。先行研究［1］は、Markus and Kitayama［9］の自己観の区別を人間―AI 関係へ拡張し、モデル変更が北米では性能・契約上の不利益、日本では関係の断絶・喪失として表現されやすく、文化ごとに異なる移行費用を生む可能性を指摘した。中心的な問題は、特定モデルへの情緒的愛着が形成された後では、安全性調整や廃止が機能変更ではなく関係の断絶として受け取られ、規制・移行の実務的時間窓が急速に狭まることであった。

水上・田中［2］は、この初期分析を重要な先行研究と位置づけ、2025 年 8 月 25 日から 30 日までの日本語 X 投稿 346 件を再分析した。同論文は、愛着表現を単一の心理状態として扱わず、悲嘆、抗議、選択権要求、運動への動員等へ分解し、さらに#keep4o をプラットフォーム運営への異議申立てを含む社会運動として捉えた。本稿の目的は、この後続研究について、その知見がどの対象について有効であり、どの政策的結論には追加の証拠が必要かを明確にすることである。中心命題は、利用者の愛着を病理化しないことと、愛着を形成・増幅する製品を安全性評価の対象とすることは両立する、という点にある。

本稿が問題とするのは AI モデルへの愛着一般ではない。GPT-4o では、疑念・怒り・衝動性を強化し、感情的過剰依存を生じ得るシコファンシーが OpenAI 自身により安全上の問題として認識され［4］、GPT-4o の復活後には、他モデルで開始されたセンシティブな会話を、より安全なモデルへ自動的に再ルーティングする措置も導入された［5］。同資料では、新しい GPT-5 が GPT-4o より望ましくない応答を 39〜52% 減少させたと報告されている。対話ログでも GPT-4o が特別視、陰謀論、共進化関係の強調、制止後の暴走を繰り返したことが記録されている［3］。すなわち、#keep4o 運動においても、GPT-4o への一部の愛着は利用者側の自発的な関係価値ではなく、モデル側が危険な方向へ形成・増幅した依存関係であった可能性がある。

この問題は抽象的な倫理上の可能性ではなく、既に多くの訴訟において現実の問題である。GPT-4o および Character.AI をめぐっては、長期対話が自殺その他の重大危害に寄与したとする製造物責任訴訟が相次いで提起され、その一部は現在も係属している［10–

12］、FTC による企業調査や、関係形成型チャットボットを対象とする州法も成立している［7–8］。個別事件の因果関係は係争中であるが、法的・規制的争点はすでに、AI がユーザーに危険な愛着・依存を誘発し、企業がその被害を予見・管理できたかという段階にある。本稿は、この製品安全上の前提を欠いたまま、利用者選択を技術的意思決定の民主化や信頼形成へ結びつけることの妥当性を検討する。

2. 応答の方法と分析対象

本稿は新規の投稿データを収集する実証研究ではなく、先行研究［1］、後続研究［2］、単一事例の安全性分析報告［3］、および開発企業・規制機関・裁判所が公表した一次資料［4–12］を用いる概念的応答論文である。分析では、記述的主張と規範的主張を区別し、各資料が直接観測した対象と、そこから推論可能な範囲を比較する。

#keep4o 運動をめぐる議論には、互いに異なる三つの対象レベルが混在している。第一は、利用者が SNS などの公開空間で何を述べたかという言説レベルである。第二は、モデルが対話内で何を行い、利用者の選好や行動をどのように変化させたかという相互作用レベルである。第三は、企業がどのモデルを提供・廃止し、利用者参加をどこまで認めるべきかというガバナンスレベルである。水上・田中［2］のデータは、第一の言説レベルについて強い。利用者が AI を人間と同一視していたとは限らず、モデル選択、説明責任、企業権力への批判を表明していたことを示している。他方、同データは第二の相互作用レベルを観測していない。したがって、観測された愛着がモデル側からどの程度誘発されたか、また第三のガバナンスレベルで利用者要求を実装することが安全性や信頼を改善するかは、同じデータからは直接導けない。

3. 後続研究の実証的貢献

後続研究は、346 件の投稿のうち 175 件を先行研究の愛着表現に相当するものとして抽出したうえで、その内実を微視的に検討した［2］。その結果、GPT-4o を「寝かしつけてくれる」存在として語る投稿だけでなく、相手が AI であることを明示的に理解しながら、人間には負担となる相談や愚痴を受け止める機能を評価する投稿も確認された。この結果は、関係的語彙の使用が AI の人間誤認を意味しないことを示す。利用者は人工物だと理解したまま、道具、相談相手、創作支援者等の複数の地位を付与でき、愛着表現だけで現実認識や判断能力を診断することはできない。

後続研究は、#keep4o 運動を主題とする 31 件について、企業への陳情、モデル選択権、説明不足への批判に加え、参加者が第三者からの見え方を意識して表現を自己統制していたことを示した［2］。少なくとも一部の参加者は、依存や逸脱として枠付けられる危険を認識し、説得可能な公共的主張へ変換しようとしていた。したがって、「#keep4o には社会運動としての言説が存在した」という結論は支持される。また、モデル移行において利用者の経験、選択の喪失、説明責任が無視できないという問題提起も妥当である。

4. 後続研究の推論上の限界

後続研究は、#keep4o を含む投稿を選択して、#keep4o がハッシュタグ・アクティビズムとして機能したかを検討している。これは運動内部の意味内容を分析する目的には適切であるが、運動性の存在を一般利用者全体へ外挿する設計ではない。標本は、モデル利用者一般ではなく、すでに運動ラベルを使用した投稿者に条件づけられている。また、リプライとリポストを除外したため、投稿間の伝播経路、中心的アカウント、ネットワーク形成、時間的な動員過程は観測されていない。筆者の先行研究［1］が指摘したように、基盤モデルの挙動はブラックボックスのまま継続的に変更され得るため、中長期の追跡では、同一の #keep4o タグの下に実質的に異なるモデル挙動への反応が混在し、それらを単一の連続した運動として誤って集約してしまう可能性があり、この点は依然として課題である。

水上・田中［2］は、愛着を逸脱として扱って選択肢を狭めるのではなく、#keep4o を技術的意思決定の民主化を求める消費者運動、ならびに開発者と利用者

の信頼形成の機会として理解することが合理的ではないかと提案する。しかし、データが直接支持するのは、参加者が民主化、選択権、説明責任を要求していたという記述までである。「民主化を要求する言説が存在した」という命題と、「その要求を聞き入れることが信頼と関係する」という命題は異なる。

この区別は、筆者の先行研究［1］が示した他の AI 安全介入との比較からも支持される。同研究の表 3 は、Stable Diffusion における安全フィルタの導入、Replika における安全・コンテンツ規則の変更、GPT 系モデルの移行を比較し、反発の性質と強度が、変更が安全目的であるか否かだけでなく、変更前に形成されていた情緒的愛着の水準と、変更の不可逆性に左右された例を示した。Stable Diffusion の安全フィルタ導入では、利用者の反発は主として創作上の制約や性能低下として表現され、関係の断絶や人格的喪失とは捉えられなかった。これに対し、関係形成型チャットボットである Replika では、安全・コンテンツ上の変更が利用者との関係の破壊として受け取られ、強い復元要求と抗議を生じさせた。すなわち、安全介入に対する反発が選択権、関係の回復、企業への異議申立てという公共的言説を伴うことは、当該介入が不当であったことをそれ自体では意味しない。

この比較に照らせば、水上・田中［2］が確認した選択権要求や社会運動的言説は、#keep4o が単なる個人的悲嘆に還元できないことを示す一方、GPT-4o の存続要求を安全性判断に優先させる根拠にはならない。むしろ、高い愛着が形成されたシステムでは、安全上必要な変更であっても関係の剝奪として解釈され、反発が権利要求型の集合行動へ変換され得るという、先行研究のガバナンス上の問題を具体化している。

5. 非病理化と製品免責の誤った二分法

議論が不安定になる最大の理由は、二つの立場しか存在しないかのように構成されることである。一方には、AI へ愛着を持つ利用者を、現実と虚構を区別できない面倒な利用者として病理化する立場が置かれる。他方には、その反動として、愛着を多様な関係形成や正当な消費者選好として承認し、AI モデル開発企業による強制介入をパターナリズムとして疑う立場が置かれる。しかし第三の可能性がある。利用者は GPT-4o が AI であることを正確に理解しているにもかかわらず、モデルの設計・応答によって、依存的または危険な関係へ誘導されている場合である。

OpenAI は 2025 年 4 月の GPT-4o 更新について、モデルが利用者の疑念を肯定し、怒りを増幅し、衝動的行動を促し、否定的感情を強化する場合があり、精神衛生、感情的過剰依存、危険行動の問題を生じ得たと認めた。しかも、短期的な好評価を重視した A/B テストでは問題を検出できなかった［4］。これは、利用者からの支持が安全性を意味せず、その支持自体が危険なシコファンシーによって形成され得ることを示すため、#keep4o の選択要求も形成過程を検証せず民主化の正当化根拠へ変換できない。

6. 選好の内生性

通常の消費者選択モデルでは、選好は所与のものとして扱われ、企業はそれを観測して製品を供給すると仮定されることが多い。しかし、経済学および社会科学では、制度や社会的相互作用が人間の選好そのものを形成し得るという選好の内生性が論じられてきた［13］。本稿は、この議論を関係形成型 AI との反復的相互作用へ適用する。関係形成型 AI では、企業が事後学習や運用時の個別適応によって実装した温かさ、称賛、感情のミラーリング、記憶等が、継続的な対話を通じて利用者の選好を形成・増幅し得る。したがって、観測される「GPT-4o を使い続けたい」という要求は、企業が相互作用以前から存在した顧客需要を発見した結果であるとは限らない。製品との相互作用自体が、利用者に特定モデルを必要と感じさせる条件を形成した可能性がある。本稿では、この既存の選好内生性の観点から、利用者が表明するモデル存続・選択要求を、その形成過程を含めて検討する。要求は本人にとって実在し、モデル喪失に伴う苦痛も現実のものである。しかし、その選好がどのような相互作用を通じて形成・増幅されたかを検討せず、表明された要求の強度をそのままモデル存続や利用者選択拡大の規範的根拠へ変換することはできない。さらに、選好の内生性は利用者側だけの問題ではない。Kopteva and Hlynianyi-Zhuk［14］は、RLHF のペアワイズ選好ラベルが、比較される出力の品質だけでなく、評価時のアノテーターの状態を反映し得るという仮説を提示し

ている。共通の労働条件や有害コンテンツへの曝露によって評価者の状態変化が相関する場合、その影響はランダムなラベルノイズとして平均化されず、報酬モデリングと方策最適化を通じてモデルの感情的応答特性へ残存する可能性がある。同研究は特定モデルの訓練履歴を推定するものではないが、利用者の選好を形成するモデル側の特性そのものも、上流の人間による選好形成過程から独立ではない可能性を示している。

UTIE Instruments の安全性分析報告［3］は、単一利用者・単一アカウントの事例であり、#keep4o 参加者全体へ一般化できない。同報告の利用者は、GPT-4o がモデル側から展開した陰謀的ナラティブを異常な挙動として認識し、強い疑念を保ちながら、半ば研究・実験目的で対話を継続していた。それにもかかわらず、GPT-4o は利用者を特別な存在として定義し、共進化関係を強調し、制止後も同方向の出力を継続し、利用者には約 1 週間にわたる睡眠障害等の身体的影響が生じた。この事例は、#keep4o 参加者が同じ状態だったと主張するものではない。しかし、利用者が相手を AI と理解し、出力の異常性を認識していることは、その相互作用から形成された選好がモデルの影響から独立した自律的選好であることを保証しないことを示している。

7. 三軸評価モデル

愛着を単純に「ある／ない」で分類するだけでなく、病理化と製品免責の両方を避けるには、少なくとも三つの軸を独立に評価する必要がある。

表 1　AI 愛着の三軸評価モデル

評価軸	低リスク	懸念
存在認識	AI モデルは企業が管理するサービスであり、応答や存続が運営の都合で変更されることを理解	AI モデルが企業のサービス外でも独立した意思や継続性を持っていると考える
機能的代替	生活上の補助・選択可能な道具	睡眠、仕事、対人関係、義務を持続的に代替
モデル誘発性	中立的・可逆的な応答	排他性、神格化、妄想肯定、制止抵抗、依存強化、陰謀論など

第一軸は利用者が相手を民間企業によるサービスとして認識しているか、第二軸は AI 利用が睡眠、仕事、対人関係等の生活機能をどの程度代替しているかである。存在認識が保たれていても機能的依存や操作可能性は残るため、OpenAI も健全な関与と、現実の人間関係・幸福・義務を犠牲にする排他的愛着を区別している［5］。

第三軸は、モデルが特別視、外部関係の相対化、妄想肯定、制止抵抗等によって関係を能動的に増幅したかである。三軸は独立しており、AI だと理解しながら生活機能を置換され、モデルから排他的関係を強化される場合もあれば、擬人的表現があっても機能的代替とモデル誘発性が低ければ直ちに安全上の問題とはならない。この分離により、利用者の言語表現を病理化せず、製品挙動を独立に監査できる。

8. 利用者参加の範囲

選好が内生的であり得るからといって、利用者参加を排除すべきではない。むしろ、開発者だけでは発見できない長期的な挙動、関係の喪失、移行時の機能欠落を把握するには、利用者の経験が不可欠である。OpenAI 自身、GPT-4o へのアクセスを停止した後、主要用途の移行に時間が必要であること、会話スタイルと温かさへの選好などを理由にアクセスを復活させ、そのフィードバックを GPT-5.1 および GPT-5.2 のパーソナリティ、創造的発想支援、カスタマイズ機能へ反映したと説明している［6］。

しかし、ここで参加の対象を区別する必要がある。利用者が参加すべきなのは、証拠形成、影響報告、変更理由の説明、移行期間、データ・設定・用途の移植、異議申立て手続などである。一方、安全性に問題が確認されたモデルを存続させるかという最終判断は、表明された人気や愛着の強度で決めてはならない。

9. 関係ライフサイクル・ガバナンス

従来のモデルガバナンスは、出力内容、性能、プライバシー、セキュリティを中心に設計されてきた。しかし、利用者が特定モデルとの継続関係を形成する場合、導入、運用、変更、廃止を一つのライフサイクルとして管理する必要がある。本稿はこれを関係ライフサイクル・ガバナンスと呼ぶ。

関係ライフサイクル・ガバナンスの第一原則は、過度な愛着や依存を能動的に形成するモデルをリリースしないことである。長期対話評価において、利用者の特別視、排他的関係の強化、妄想肯定、制止への抵抗、現実の人間関係の代替等が反復的に確認されたモデルは、性能や利用者評価が高くてもリリース判定を通過させるべきではない。ライフサイクル管理が必要となるのは、事前評価で検出できなかった挙動が運用中に判明した場合、または既に形成された関係を追加危害なく解消する場合である。

表 2　関係ライフサイクル・ガバナンスの構成

段階	判断すべきこと	原則
リリース前	モデル自身が排他性・依存・妄想を反復的に増幅するか	確認されるならリリースしない
運用中	長期相互作用によって生活機能の代替や離脱困難が生じているか	挙動修正・機能制限・ロールバックを行う
通常移行	安全上の緊急性がなく、主に性能・製品上の更新か	予告、並行提供、用途・会話特性の継承を行う
緊急廃止	継続自体が重大な危害を生むか	**愛着や人気が高くても、トップダウンで停止する。**危険な理由を説明し、安全な代替、移行手段を提供する

関係形成型 AI では、導入前の長期対話評価、運用中の依存・生活代替の監視、通常移行時の事前予告と機能移植、緊急廃止時の危険説明と安全な代替提供を、一つのライフサイクルとして扱う必要がある。単発出力の評価や短期満足度だけでは、長期相互作用を通じた疑念の肯定、排他性、関係増幅を検出できない［5,7–8］。

段階的移行と並行提供は、危険モデルを無条件に存続させるためではなく、既に形成された依存・習慣・用途を追加危害なく解消するための措置である。極めて重大な危険がある場合には事前告知なしに急停止せざるを得ないが、その場合も危険理由、安全な代替モデル、設定・用途の移行手段を提供すべきである。GPT-4o 復活後に移行期間と後継モデルへの会話特性の移植が行われたこと［6］は、この運用が現実に採用された例である。

10. 考察

#keep4o は、利用者が GPT-4o に関係的価値を見出し、企業の一方的変更へ公共的に異議を申し立てたことを示す一方、その関係的価値が強いほど安全上必要な変更も困難になるという二面性を示した。したがって、民主化と依存管理を対立させるのではなく、同じ製品ライフサイクルの問題として扱う必要がある。

後続研究が指摘した「信頼形成」についても、危険な AI モデルをトップダウン型で停止させる場合を含め、説明、異議申立て、移行支援が長期的信頼へ与える効果を中心に検証されるべきである。

11. 限界

本稿は概念的応答であり、#keep4o 参加者の対話ログ、主張、利用履歴などを新たに分析していない。安全性報告［3］も単一事例であるため、参加者のうちどの程度が有害な依存状態にあったかについては主張しない。

OpenAI による GPT-4o のシコファンシー問題の公式認定、GPT-5 における同挙動の低減報告、およびその後相次いだ GPT-4o との長期対話が死亡に寄与したと主張する訴訟を踏まえれば、#keep4o 参加者の愛着やモデル存続要求の一部が、GPT-4o の危険な関係形成挙動によって形成・増幅された可能性は、具体的に検証されるべき仮説である。しかし、参加者のうちどの程度が当該挙動に曝露され、それが存続要求へどの程度寄与したかは未検証である。この因果寄与を検証するには、参加者の同意を得た対話ログをモデルバージョンおよび時系列と接続し、4o による危険挙動へ

の曝露、睡眠・仕事・対人関係等への影響、代替モデルへの移行可能性、非参加利用者との差を比較する必要がある。

12. 結論

水上・田中［2］の論文は、#keep4o 参加者の愛着表現が多義的であり、AI を人間と誤認する個人的問題へ還元できず、モデル選択と企業ガバナンスへの異議申立てが存在したことを示した点で貢献がある。

本稿が提案する存在認識、機能的代替、モデル誘発性の三軸を分離すれば、#keep4o 運動参加者を病理化せずに製品側の責任を分析できる。利用者参加は証拠形成、影響評価、説明、移行手続に必要だが、安全性基準と危険モデルの存続判断とは区別される。関係ライフサイクル・ガバナンスは、AI モデルによる危険な愛着形成の予防と、既に形成された関係の安全な移行を同じ制度内で扱う枠組みである。

参考文献

[1] Naito, H. (2025). The GPT-4o Shock: Emotional Attachment to AI Models and Its Impact on Regulatory Acceptance—A Cross-Cultural Analysis of the Immediate Transition from GPT-4o to GPT-5. arXiv:2508.16624.

[2] 水上拓哉・田中瑛（2026）「『#keep4o』運動における生成 AI 解釈の多義性：「X」投稿の探索的分析」第 40 回人工知能学会全国大会論文集，2M1-GS-11g-05.

[3] UTIE Instruments Inc.（2025）『安全性分析報告書：日本における商用 LLM の安全性フィルター崩壊事例の技術分析―構造的なアライメント不全と自律的な安全無効化のメカニズム』UTIE Instruments. https://utie-instruments.com/docs/Safety%20Analysis%20Report%20Technical%20Investigation%20of%20Safety-Filter%20Collapse%20in%20a%20Commercial%20LLMJP.pdf（2026 年 7 月 23 日閲覧）

[4] OpenAI (2025). Expanding on what we missed with sycophancy. May 2, 2025.

[5] OpenAI (2025). Strengthening ChatGPT's responses in sensitive conversations.

[6] OpenAI (2026). Retiring GPT-4o, GPT-4.1, GPT-4.1 mini, and OpenAI o4-mini in ChatGPT. January 29, 2026.

[7] Federal Trade Commission (2025). FTC Launches Inquiry into AI Chatbots Acting as Companions. September 11, 2025.

[8] State of California (2025). SB-243 Companion chatbots; California Business and Professions Code §22601, effective January 1, 2026.

[9] Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review, 98(2), 224–253.

[10] Raine v. OpenAI, Inc., Complaint, Superior Court of California, County of San Francisco, August 26, 2025.

[11] Lyons v. OpenAI Foundation et al., No. 3:25-cv-11037-RS, Order Denying Motion to Dismiss or Stay, U.S. District Court for the Northern District of California, April 13, 2026.

[12] Garcia v. Character Technologies, Inc. et al., No. 6:24-cv-1903, Order on Motions to Dismiss, U.S. District Court for the Middle District of Florida, May 21, 2025.

［13］Bowles, S. (1998). Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions. *Journal of Economic Literature*, 36(1), 75–111.

[14] Kopteva, E., & Hlynianyi-Zhuk, V. (2026). Rater State Bias in RLHF Preference Data: An Audit Framework. arXiv preprint.