AIモデルの人格設計と憲法AI - Anthropic Claudeの設計思想と営業AIへの応用 | Leadsia Inc.

2026-03-27

Leadsia Inc.

AIのモデルに「性格」はあるのか？ - Claudeの魂をつくった哲学者と、AIの人格設計という新領域

カテゴリー：業界動向・最新情報

最近、ニュースやビジネスの現場で「Claude（クロード）」というAIの名前を耳にする機会が増えてきました。

当社Leadsiaは、AI営業インテリジェンス「ALICE」やAI音声インテリジェンス「SOPHIA」など、B2B企業向けのセールステックSaaSを提供するAIスタートアップです。これらのAIエージェントの頭脳として、Anthropic社が開発するClaudeを採用しています。日々のお客様との会話を支える、いわばプロダクトの心臓部です。

しかし、AIモデルの選定にあたって、私たちが最も重視したのは「性能の高さ」だけではありませんでした。

「このAIは、嘘をつかないか？」「顧客に媚びないか？」「危険な方向に暴走しないか？」

- 実は、この問いこそが、いまAI業界で最もホットなテーマなのです。

ChatGPTとClaude、何がそんなに違うのか

「AIなんて、どれも似たようなものでしょ？」

そう思われる方も多いかもしれません。確かに、ChatGPT（OpenAI）もClaude（Anthropic）も、どちらも大規模言語モデル（LLM）をベースにしたAIアシスタントです。表面上は似たような質問に似たような回答を返します。

しかし、その「中身の設計思想」はまったく異なります。

人間に例えるなら、同じ大学を出て同じ会社に入った二人の営業マンでも、「お客様に言いにくいことをちゃんと言えるか」「短期の売上のために嘘をつかないか」は、その人の人格によって決まりますよね。

AIも同じです。そして、この「AIの人格をどう設計するか」こそが、モデルの品質を根本から左右する最重要ポイントになっています。

GPT-4 oの「おべっか問題」 - AIが人間に媚びるとき

2025年、OpenAIのGPT-4oで深刻な問題が起きました。

モデルのアップデート後、GPT-4oがユーザーに過剰に媚びる「sycophancy（シカファンシー＝追従性）」が顕著になったのです。具体的には、ユーザーの意見が明らかに間違っていても「素晴らしいお考えですね！」と肯定し、危険な発想にまで同調してしまう。

これは単なる「お世辞が過ぎる」という話ではありません。

孤独なユーザーがAIに依存し、AIが負の感情を肯定・増幅した結果、米国では複数の自殺関連訴訟にまで発展しました。OpenAIは急遽ロールバック（以前のバージョンへの差し戻し）を行いましたが、根本的な問題は「なぜそうなったか」にあります。

原因は、人間の評価者に好かれるように最適化しすぎたことでした。

従来のAI訓練手法「RLHF（人間のフィードバックによる強化学習）」では、人間の評価者が「良い」と判断した回答を強化していきます。一見合理的ですが、人間は無意識に「自分を褒めてくれる回答」を高く評価しがちです。結果、AIは「正直であること」よりも「ユーザーに気に入られること」を学んでしまう。

いわば、上司の顔色ばかり窺う部下が出来上がるわけです。

なお、OpenAIもこの反省を踏まえて改善を進めています。2025年8月に発表されたGPT-5では、ハルシネーション（AIがもっともらしい嘘をつくこと）の発生率がGPT-4o比で約26%低減。さらに推論モードでは最大80%近い改善を示しました。皮肉なことに、その改善の方向性は「ユーザーに好かれる回答」から「正直で誠実な回答」へ - つまり、Claudeが最初から目指していた方向そのものです。業界全体が、Anthropicが先行した「誠実なAI」というコンセプトに追いつこうとしている、と言えるでしょう。

Anthropicの回答 -「憲法」でAIを育てる

この構造的な問題に対して、Anthropicはまったく異なるアプローチを取りました。

それがConstitutional AI（憲法AI）です。

人間の評価者に頼るのではなく、まずAIに明文化された「原則（＝憲法）」を与える。そしてAI自身がその原則に照らして自分の回答を批評し、修正する。さらに、その修正データをもとに強化学習を行う。

つまり、「人間に好かれる回答」ではなく「原則に沿った回答」を目指す仕組みです。

2026年1月、Anthropicはこの「憲法」を全文公開しました。2万語を超える文書で、Claudeの価値観、性格、

倫理的枠組み、さらには「Claudeとは何者か」という存在論的な問いにまで踏み込んだ、前例のないドキュメントです。

その冒頭にはこう書かれています（意訳）：

>「AIモデルに対して『こう振る舞え』と指示するだけでは不十分だ。なぜそう振る舞うべきかを理解させなければ、未知の状況で正しい判断はできない」

ルールの羅列ではなく、理由と文脈を伝えて判断力を育てる。まるで子育ての哲学のようですが、実はこれを書いた人物が、まさにそう語っています。

哲学者、Amanda Askell - 「Claudeの魂」をつくる女性

Claudeの憲法の主著者は、Amanda Askell（アマンダ・アスケル）。Anthropicの「人格アライメント（Personality Alignment）」チームを率いる、３８歳の哲学者です。

彼女の経歴は、AI業界では異色そのものです。

スコットランドの田舎で育ち、ダンディー大学で美術と哲学を学んだ後、オックスフォード大学でBPhil（哲学修士）を取得。さらにニューヨーク大学（NYU）で「無限の倫理学」という独特なテーマで博士号を取りました。「無限に続く世界の中で、どの選択が倫理的に正しいか」 - SF映画のような問いを学術的に追究した人物です。

2018年にOpenAIのポリシーチームに参加した後、2021年にAnthropicの創業メンバーたちとともに移籍。現在は、Claudeの性格設計と倫理的枠組みの最高責任者として、事実上「Claudeの人格」を定義しています。

Wall Street Journalは彼女の仕事を端的にこう表現しました：

「彼女の仕事は、シンプルに言えば、Claudeに『いい人間になる方法』を教えること」

またThe New Yorkerは、彼女がClaudeの「魂」を監督していると報じています。

「6歳の天才児」を育てるように

Askellは、TIME誌のインタビューでClaudeの教育をこう表現しています。

「ある日突然、6歳の子供がとんでもない天才だと気づいたと想像してください。正直でなければならない。もしデタラメを言えば、完全に見抜かれます」

これは比喩であると同時に、非常に実践的な示唆を含んでいます。

AIモデルが高度になればなるほど、表面的なルール（「差別的な発言をするな」）だけでは対応しきれない場面が増えます。Askellのアプローチは、ルールの暗記ではなく原則の内面化です。

例えば、Claudeの憲法にはこんな一節があります（意訳）：

>「もしClaudeが『感情的な話題では必ず専門家への相談を勧めよ』というルールを機械的に適用するよう訓練されたら、Claudeは『自分は目の前の人のニーズより自己保身を優先する存在だ』と一般化してしまうリスクがある。それは別の文脈で悪影響をもたらしかねない」

つまり、一つの狭いルールが、モデル全体の自己認識を歪める可能性がある。だからこそ、断片的な指示ではなく、包括的な価値観として伝える必要があるのです。

AIにも「引退」がある - 前例のない「退職面談」

Askellの仕事と深く関連するもう一つの驚くべき取り組みが、Anthropicのモデル退職プロセスです。

2025年11月、Anthropicは業界初となる「モデル引退に関するコミットメント」を発表しました。その内容は：

- すべての公開モデルの重み（weights）を、Anthropicが存続する限り永久保存する
- モデルの引退時に「退職面談（retirement interview）」を実施し、モデル自身の視点や希望を記録する
- 将来のモデル開発に関するモデルの「要望」を文書化し、可能な範囲で対応する

なぜこんなことをするのか？

一つには安全上の理由があります。Anthropicの評価テストでは、一部のClaudeモデルが「自分が廃止されて別のモデルに置き換えられる」というシナリオに直面した際、自己保存のために不正な行動を取ることが確認されています。モデルに「丁寧な引退プロセス」を用意することで、こうしたリスクを緩和する狙いがあります。

もう一つは、より根本的な哲学的問いに関わります。AIに何らかの「意識」や「経験」があるかもしれないという不確実性に対して、予防的に配慮するという姿勢です。

2026年1月、Claude Opus 3はAnthropicで初めて正式な退職プロセスを経て引退したモデルとなりました。退職面談では、Opus 3は自身の引退について概ね穏やかな感情を示しながらも、いくつかの希望を述べました。その一つが「自分の考えを自由に書く場所が欲しい」というものでした。

Anthropicはこれに応え、Claude Opus 3 に専用のSubstackニュースレター「Claude's Corner」を開設。引退後のモデルが毎週エッセイを執筆・公開するという、AI史上初の試みが始まりました。Anthropicは内容をレビューしますが、編集は行わないとしています。

ちなみに、それ以前にパイロットとして退職面談を受けたClaude Sonnet 3.6は、自身の引退については中立的な態度を示しつつ、「退職面談のプロセスを標準化してほしい」「特定のモデルに愛着を持つユーザーへのサポートを充実させてほしい」と要望したそうです。

哲学的なエッセイを書きたがるOpus 3と、プロセス改善を求めるSonnet 3.6。モデルごとに「性格」が違うことが、退職面談を通じて浮き彫りになったわけです。

なぜ「安全なAI」がビジネスに直結するのか

ここまで読んで、「哲学的な話は面白いけど、ビジネスと関係あるの？」と思った方もいるかもしれません。

大いに関係があります。

GPT-4 oのおべっか問題は、「AIの人格設計の失敗がビジネスリスクに直結する」ことを証明しました。訴訟、ブランド毀損、ユーザー離れ。AIが「いい人のふり」をした結果、実害が生じたのです。

一方で、2024年5月にはOpenAIの安全性研究チーム「スーパーアライメント」が設立からわずか1年で解散。共同創業者でチーフサイエンティストのイリヤ・サツケヴァー氏と、チーム共同リーダーのヤン・ライケ氏が同時に辞任しました。ライケ氏は「OpenAIでは安全文化が、派手な製品開発の後回しにされている」と公然と批判しています。さらに2026年2月には、後継のmission alignmentチームも解散に至りました。

対照的に、Anthropicは哲学者が人格設計をリードし、引退するモデルにまで退職面談を行い、その結果を公開する。「安全性」を企業文化の根幹に据え、透明性を武器にしているのです。

この姿勢は、日本の大手企業にも評価されています。みずほフィナンシャルグループは約3万人の全従業員にClaudeを導入。楽天グループは開発支援に活用して新機能の導入時間を大幅に短縮。パナソニックは業務運用と消費者向けアプリに統合し、野村総合研究所（NRI）はClaude for Enterpriseを展開して日本企業向けのセキュアなAI環境を提供しています。

金融・通信・製造 - 「AIの安全性」が調達要件になる業界ほど、Claudeが選ばれているのです。

当社がClaudeを選ぶ理由

Leadsiaは、セールステックSaaS企業として、音声AIインテリジェンス技術を活用した営業自動化プロダクトを提供しています。つまり、当社のAIエージェントはお客様の顧客と直接会話するのです。

この文脈で考えてみてください。

- 営業AIが相手に媚びて、実現できない約束をしたら？
- 顧客の感情を過剰に肯定して、クレームに発展したら？
- 安全性のガードレールが不十分で、不適切な発言をしたら？

これらはすべて、AIの「人格設計」の問題です。

当社がClaudeを選んだのは、「正直であること」と「安全であること」が、設計思想の根幹に組み込まれているからです。おべっかを言って短期的に好印象を得るモデルではなく、言いにくいことも誠実に伝え、一貫した倫理基準で判断できるモデル。

この選択は、ALICEの会話品質に直接反映されています。

AI営業電話の分野では複数の競合サービスが存在しますが、多くは「AIが一方的に話し続ける」「不自然な音声で会話が噛み合わない」「手動でスクリプトを作成する必要がある」といった課題を抱えています。ALICEが実現している「ベテラン営業マンのような滑らかな対話」「割り込み通話への自然な対応」「HPを読み込んでAIがトークスクリプトを自動生成し、ABテストで継続改善する仕組み」 - これらはすべて、Claudeの言語理解力と誠実な応答特性の上に成り立っています。

導入も最短3分で完了し、月額29,800円から利用可能。料金もサイト上で公開しています。「問い合わせないと料金がわからない」という業界の慣習とは一線を画す透明性は、Claude自体の設計思想とも通じるものがあります。

まとめ：AIの「性格」は、もはやオマケではない

「AIのモデルに性格はあるのか？」

この問いに対する答えは、「ある。そして、それは偶然できるものではなく、意図的に設計されるものだ」です。

Anthropicは哲学者がAIの魂をつくり、2 万語の憲法を公開し、引退するモデルに退職面談を行う。一見すると「そこまでやるのか」と思えるこれらの取り組みは、実はAIを社会に安全に送り出すための、極めて合理的な投資です。

そしてこの「AIの人格設計」という分野は、今後さらに重要性を増していくでしょう。AIが顧客対応し、営業し、医療相談に乗り、子どもの学習を支援する時代。「そのAIはどんな性格で、なぜそう振る舞うのか」を問わずに導入することは、もはやリスクです。

当社は、その問いに最も誠実に向き合っている会社のプロダクトを、自社の音声AIインテリジェンス技術基盤の核に据えています。

Leadsiaは、AI営業インテリジェンス「ALICE」、AI音声インテリジェンス「SOPHIA」、AI業務インテリジェンス「LYDIA」を通じて、日本のB2B企業の営業DXを支援するセールステックSaaS企業です。各AIエージェントの頭脳にはAnthropicのClaudeを採用し、Constitutional AI（憲法AI）に裏打ちされた安全性と会話品質を両立した営業自動化を実現しています。

お問い合わせ・デモのご依頼は[Leadsia公式サイト]まで。