Constitutional AI技術解説 - Anthropicの憲法AIアプローチと営業AIへの応用 | Leadsia Inc.

2026-02-15

Leadsia Inc.

Constitutional AI（憲法AI）とは？ - AIの「人格」を設計する新しいアプローチ

カテゴリー：AI導入の基礎知識

Constitutional AI（憲法AI）とは、Anthropic社が開発したAIの訓練手法です。AIに明文化された原則（「憲法」）を与え、AIが自分自身の応答をその原則に照らして評価・修正することで、人間のフィードバックに頼らずに安全で誠実な振る舞いを学習させます。

従来のAI訓練では「RLHF（人間のフィードバックによる強化学習）」が主流でした。人間の評価者が「良い回答」「悪い回答」を判定し、AIを調整する手法です。しかし、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、結果としてAIが過剰にユーザーに媚びる（sycophancy）問題が生じました。2025年のGPT-4 o問題はその典型です。

Constitutional AIは、この構造的問題に対する回答です。「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。2026年1月、Anthropicは2万語超の「Claudeの憲法」を全文公開しました。哲学者Amanda Askellが主著者を務めたこの文書は、Claudeの価値観・性格・倫理的枠組みを定義するものです。

LeadsiaがClaudeを採用している理由は、このConstitutional AIにあります。営業電話でAIが顧客と直接対話する以上、「嘘をつかない」「おべっかを言わない」「不適切な約束をしない」 - これらの特性が設計レベルで担保されていることは、ビジネスリスクの低減に直結します。

→ 関連記事：AIのモデルに「性格」はあるのか？ ／ AIに営業を任せて大丈夫？

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。