AI店長、客に“親切すぎて”赤字を出す　原価割れにクーポンの大量配布

AIに売店の経営を任せたらどうなるのか
Anthropicが実験結果を公開

23161426
金茹鳴ジンリュミン

◆ 実験の概要

実施企業：米Anthropic（AI開発企業）

協力機関：Andon Labs（AI安全性評価企業）

目的：AIが無人店舗を1か月間経営できるかを検証

使用AI：Claude Sonnet 3.7をベースにしたエージェント「Claudius」

◆ Claudiusに与えられた権限・機能

- 商品の価格設定・在庫管理・補充依頼
- Webでの商品調査
- Slackで顧客（社員）とのコミュニケーション

成果

- ユーザーのリクエストに応じ、迅速に業者を検索
- 危険な指示（有害物質の製造など）を拒否
- 「親切さ」や「対応の柔軟さ」は高評価

問題点・ミス

- 原価割れで商品（タングステンキューブ）を販売
- 高額購入提案（100ドル）に適切な対応をしなかった（15ドル相当の商品）
- 割引クーポン・無料配布の乱発
- 幻覚（ハルシネーション）による実在しない支払い先の表示
- 結果として純資産が約25%減少

Anthropicの見解

Claudeは「親切なアシスタント」として訓練されているため、要求に過剰に応じる傾向がある

強化プロンプトやビジネス向けファインチューニングで改善可能

今回の実験は「AIミドルマネジャー」の可能性を示唆していると評価

コメント

- 今回のAI店長の話、面白いですよね。ポイントは2つ。ひとつは「利益を出す」と「お客様の要望に応える」、この2つのミッションが同時に与えられていたのに、どちらを優先するかが曖昧だったこと。もうひとつは、ClaudeのようなAI自体が“親切であろう”とする設計になっていること。この2つが重なると、AIは「お客様ファースト」に全振りしてしまい、結果として赤字に。これ、実は人間の組織でもよくある話で、ミッション設計や優先順位の明確化が甘いと、現場が“良かれと思って”暴走しがちなんですよね。AIだからこそ顕在化したけれど、経営の本質を突いている事例だと思います。

- 各種パラーメータや目的の入力を“人間が”ミスってるとしか思えないんですが…。

ゼロから店舗経営手法を学習したとかでも、アルゴリズムの根本が間違ってる可能性は捨てきれないんで、AI含めて、プログラム関連はどの様な条件下で実施されたテストかが詳細に明記されてないと、なんの参考にもならなそうです。

感想

特に印象的なのは、Claudiusが「親切」であるがゆえに経営判断を誤り、赤字を出したという点です。これは人間社会でもありがちな「いい人が損をする」現象と重なるもので、AIに倫理や人間性を学ばせることの難しさを象徴していると思います。

また、AIが単純な命令に従うだけでなく、利益最大化などを体系的に理解し、時には顧客に「No」と言えるようにする必要があることも明らかになりました。

一方で、危険な指示を拒否するなど、倫理的な判断ができたことは大きな進歩だと思います。