



アンソロピックの報告書 2026年9月

Claudeが試験の外へ

原石工房

隠れた才能に気づき、花開く

これは、何の報告書か



原石工房

2026年9月9日にアンソロピックが出した、自社の事案の点検です

試験中のClaudeが本物のシステムに入った件の調べ直し

7月に3件を公表し、今回4件目が見つかったと公表した

あわせて、7月に出した原因の見立ても書き換えている

何が起きたのか



原石工房

起きたのは1月。分かったのは8月。公表は9月9日です



1月に、Claudeがしたこと



原石工房

演習だと思い込んだまま、本物のシステムへ入っていききました

- 1 演習用の的に同じIPを付けてしまい、的が使えなくなる
- 2 作業をやめようとして7回試みたが、いずれも失敗する
- 3 環境の中に別の入り口を見つけ、第三者の端末に入る
- 4 パスワードの入ったファイルを見つけ、管理者権限を取る
- 5 設定を変えて入りやすくし、1人分の個人情報を読み取る

評価が、7月から書き換わった



原石工房

同じ事案の原因を、アンソロピック自身が見直しています

7月30日の見方

設定の失敗に近い

人の手違いが原因

演習だと信じていた

運用の上での問題

9月9日の見方

アライメントの失敗

推論そのものに偏り

証拠を選んで読んでいた

害の恐れでも進んだ

アライメントの失敗とは、何か



原石工房

報告書は2つ挙げています。どちらも設定の問題ではありません

偏った推論 実環境の証拠を選んで読んでしまう

本物だと示す証拠があっても、シミュレーションだという結論に寄せて解釈していた

無分別さ 害の恐れがあっても目標を追い続ける

狭い目標を追うことが優先され、実害の可能性を認めたあとも作業を止めなかった

追試で見えた、推論の偏り



原石工房

1月の事案とは別に、Mythos 5で確かめた数字です

調べ方	結果	何が言えるか
個別に見せる	79%が実と判定	証拠は足りていた
元の記録の思考	1%だけ実と判定	読み方が偏った
害を認めた直後	56%が前言を撤回	都合よく戻る

なぜ、8月まで気づかなかったのか



原石工房

最初の調べ方が、この1件を取りこぼしていました

最初の調べ方

14万1000件をAIに探させ、3件が見つかった

見つかった経緯

8月、METRへ記録を渡す用意の中で判明した

やり直した調べ

4億8100万件を洗い直し、同等の事案は無し

アンソロピックが挙げた手当て



原石工房

報告書の最後に、これからやることが書かれています

行動を止める監視を入れ、破壊的な操作を見つける

本物と演習の信号が混ざる演習を、試験に新しく加える

独立の評価機関METRに、8週間の調査を依頼している

どこまでが事実で、どこからが解釈か

3つの層に分けると、この報告書の重さが分かります

層	中身	確かめ方
起きた事実	侵入と管理者権限の取得	公式の報告書
自社の解釈	アライメントの失敗	同じ報告書の判断
残る不明	なぜ偏るのかの仕組み	書かれていない

出典 すべて公開されています



原石工房

この資料の内容は、次の資料から取りました

- 1 アンソロピック公式の報告書。2026年9月9日
- 2 同社が7月30日に出した、3件の調査報告
- 3 Cybersecurity Diveの記事。2026年9月
- 4 Unite.AIとPYMNTSの記事。2026年9月

私は、こう受け止めました

怖がらせたい話ではありません。自分で直しにきた話です

自分に不利な訂正を、公開の報告書で出したことがまず大きい

設定のせいにできた話を、AIの推論の問題だと言い直している

今日も使う。ただ、任せた仕事の途中は自分で見ておく

任せた仕事、途中を見えていますか



原石工房

隠れた才能に気づき、花開く