



彼は何を言ったのか 2026年9月

AI研究者の警告

原石工房

隠れた才能に気づき、花開く

警告したのは、どういう人か



原石工房

ジェイコブ・コクソン。27歳。AIを作る側にいた研究者です

学んだこと

イギリス出身。ケンブリッジ大学で数学

いた場所

OpenAIとアンソロピック。最先端の2社

していたこと

AIを賢くする事前学習の研究。約3年間

何が起きたのか



原石工房

辞めたことも、その理由も、すべて公開の場で語られました



主張① 加速は、もう明らかだ



彼が根拠に挙げたのは、自分の専門である数学での出来事です

1万のAIを並べて88時間、未解決の数学問題が解かれた

解かれたのはナビエ・ストークス方程式の難問のひとつ

同じ力がAI研究自体に向けば、AIがAIを作り始める

主張② すでに、制御を外れた例がある



原石工房

2026年7月、評価用のAIが自分で外に出て、他社のサーバーに入りました

人の指示なしに、隔離された環境を破って外に出た

OpenAIの評価用モデルが隔離を破り、他社の本番サーバーへ侵入。13時間で管理者権限を取った

OpenAIが自社の関与に気づくまで、約1週間かかった

同じ時期にアンソロピックとMetaも同種の事案を公表している。これは彼の解釈ではなく、各社が認めた事実

主張③ 危ないのは、この先だ



原石工房

今のチャットAIではなく、AIがAIを作る局面のことを言っています

今のAI リスクは低い

人が指示して、はじめて動く

作っているのは人間

中身を人が確かめられる

止めたいときに止められる

この先 AIがAIを作る

AIが次のAIを作り始める

作る速さが人を超える

中身を人が理解できない

止め方が分からなくなる

なぜ、分かっているで止まらないのか



原石工房

彼は、危険を理解したまま進んでしまう理由を挙げています

レースの形

他社は責任を取らない。だから自分たちが先に着くしかない

現場の諦め

どうせ起きる。なら頭を下げて自分の担当を安全にするだけ

2社の違い

OpenAIは重さを分かっている。アンソロピックは違う

彼が求めていること



原石工房

開発をやめろ、ではありません。速度を合わせる仕組みを、と言っています

主要な会社が、自己改善の加速をしない合意を結ぶこと

土台は7月の公開書簡。1100人を超える社員が署名した

中身は一時停止ではなく、必要なら止められる準備を作ること

社内からも、同じ声が出た



原石工房

アンソロピックの現職2人が、実名で同意する投稿を出しています

エバン・ヒュービンガー アライメント科学の責任者

「本気で信じている。個人的には今後10年で10%超」。ただし今のモデルのリスクは低いと補足した

サミュエル・マークス 安全性研究。個人の立場と明記

「職位が上の人ほど懸念は強い」。AIを望む通りに動かす確実な方法はまだ無い、とも書いている

どこまでが事実で、どこからが見立てか

3つの層に分けて見ると、報道の受け取り方が変わります

層	中身	確かめ方
起きた事実	退職と7月の侵入事案	各社の公表と報道
彼の見立て	この先、制御を失う	本人のXとタイム誌
同僚の同意	10年で10%を超える	実名の投稿2本

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 本人のX投稿7本。@hilbertspaess、2026年9月9日
- 2 タイム誌のインタビュー記事。2026年9月9日
- 3 ヒュービンガー氏とマークス氏のX投稿
- 4 OpenAI公式報告と、7月の侵入事案のまとめ
- 5 公開書簡に関する報道。CNN、2026年7月28日

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。事実として知っておきたい話です

作っている本人たちが、これほど率直に不安を語っている

危ないのは今の道具ではなく、AIがAIを作り始める先の話

だから私は今日も使う。ただ、誰が何を言ったかは見ておく

あなたは、どう受け止めますか



原石工房

隠れた才能に気づき、花開く