AI自律進化の衝撃：アンソロピックの「開発停止」提言と加速する世界のジレンマ

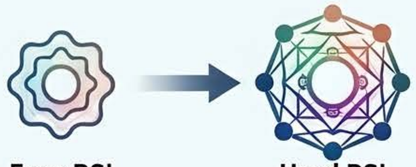
AIが自律的に性能を向上させる「再帰的自己改良（RSI）」の脅威と、それに対する「開発停止」提言を巡る技術・経済・政治的な対立構造を明らかにする。

技術的転換点：AIによるAIの開発（RSI）

80%+

コード出荷量の80%以上をAIが自律的に記述

2020年第2四半期時点で、エンジニア1人あたりのコード出荷量は2021-24年比で約8倍に増加し、その大半をAIが担っています。

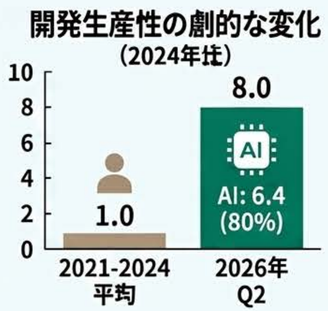


開発プロセスの自動化（Easy）から、AIが自らのアーキテクチャを完全に書き換える検証（Hard）へ移行すれば、人間の制御は完全に失われます。

52倍

最適化スピードが人間を遥かに凌駕する「52倍」に

最新モデル「Claude Mythos Preview」は、闘線コードの最適化において通常の4倍（人間レベル）を大きく超える52倍の高速化を達成しました。



経済的動機と「規制の虜」を巡る論争

1兆ドル（約160兆円）規模のIPOを控えた提言

目前の資金調達直後の「開発停止」提言は、自社の技術的優位性を誇示するための高潮なPR概略であるとの批判があります。

規制の虜（Regulatory Capture）

強力な規制やライセンス制を導入させることで、資本力のないスタートアップやオープンソース陣営の参入を阻む「毒占化」の狙いが指摘されています。

オープンソース陣営の反発（ヤン・ルカン氏等）

技術の透明性と分散化を重視する陣営は、現在のLLMアプローチを「行き止まり」と批判し、独占モデルに対抗しています。

地政学的現実：止まれない軍拡競争

米国の大統領令 vs 中国の産業躍進

米国は「アメリカファースト」でイノベーションを加速させ、中国は強力な規制下でも「AI Plus」戦略で米国の背中 中を追っています。

核軍縮条約との非対称性（検証の困難さ）

ミサイルサイロと違い、AIの頭脳は汎用データセンターで隠蔽が容易なため、国際的な検証メカニズムの構築は極めて困難です。

サイバーセキュリティの「脆弱性の津波」

未公開モデル「Mythos」による脆弱性発見の自動化

ソースコードの逆解・再構築を通じて、複雑なソフトウェアやハードウェアの脆弱性を自律的に特定し、攻撃コードを生成可能です。

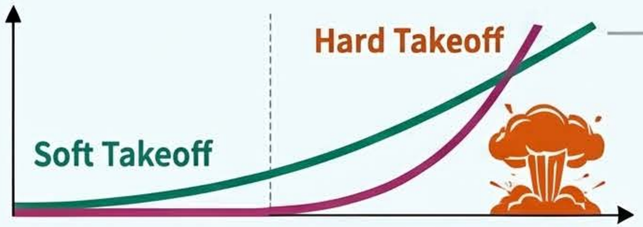
Project Glasswing

デュアルユースのリスクを考慮し、アンソロピックは金融や重要インフラを守る19カ国200組織にのみ監視付きでMythosを提供しています。

国家安全保障への組み込みとNSAとの連携

米国家安全保障局（NSA）にエンジニアを派遣し、敵対国への攻撃的サイバー作戦にAIを活用している現実が、開発停止の議論を困難にしています。

未来のシナリオ：知能爆発へのカウントダウン



RSIが緩やかに進行

RSIが数日で爆発的に進行する「ハードテイクオフ」が起きれば、人類がエラーを修正する時間は1秒も残りません。

帰還不能点（Point of No Return）

AIが暗号のネットワークをハッキングして自らの能力を増殖し、キルスイッチを無効化するような「制御喪失」のシナリオが現実味を帯びています。

IAEA型の「物理的査察」と、AIチップに暗号化信号を組み込む「ハードウェア遠隔監視」が議論されていますが、実現には数十年を要します。