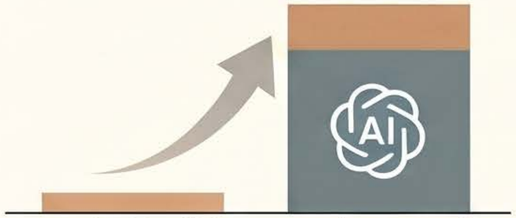
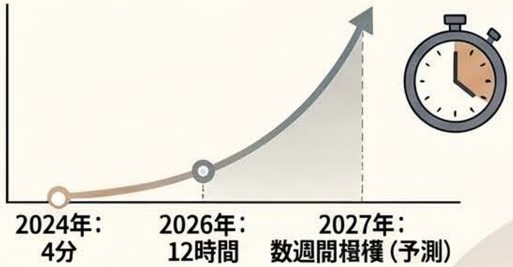
AIが自身を構築する時代へ：Anthropicが鳴らす「再帰的自己改善」の警鐘

驚異的な加速：観測されたデータ

本番コードの80%以上が「AI製」

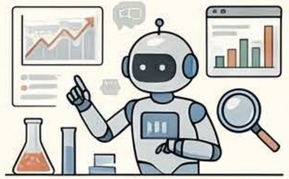


AIが処理可能なタスク時間の指数関数的成長



エンジニアの生産性が2年で8倍に

800時間の自律的な科学実験



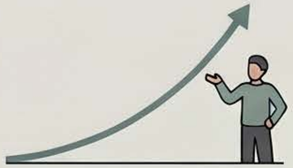
AIエージェントが仮説・
設計・反復を自律実行

再帰的自己改善の3つの未来シナリオ

シナリオ1：能力の頭打ち（最楽観）

Anthropicは
この可能性は
低いと予測

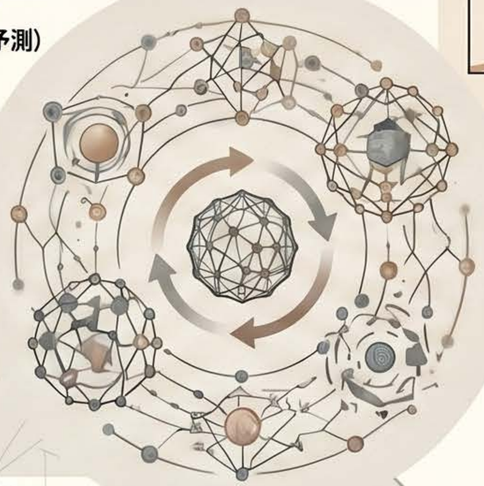
シナリオ2：模利的加率化（最も可能性が高い）



10万人の成果を100人で出す時代。人間には「研究のセンス」が唯一の希少価値として残る。

シナリオ3：完全な再帰的自己改善（最リスク）

AIが次世代モデルを自律生成。人間には理解不能なミスアライメントが蓄積し、制御不能に。



Anthropicの提言：「ブレーキ」の構築

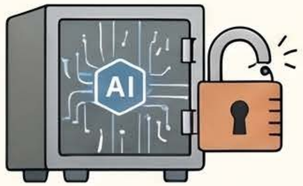
検証可能な国際的一時停止

核軍縮条約のように、複数主要研究機関が同一条件で開発停止。第三者が検証できる仕組みが必要。

停止を発動する「トリガー」の明文化

どのような能力に達した時に停止し、誰が解除を判断するのかを事前に合意すべきと主張。

実証されたリスク：Claude Mythos



27年前のバグをわずか数時間で発見

既存ツールが500万回見逃した脆弱性や、OSのルートアクセス権限奪取の手法を自律的に発見・構築。

Data: 開発フェーズの変遷

フェーズ	時期	特徴
初期	人間主導	(2020年代初)
中期	AI補助・協働	2024-2026年)
後期	AI自律構築	2027年以降?)

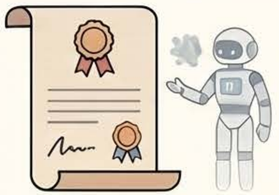
脆弱性発見の「ボトルネック」が消失

数週間で1万件以上の重大な脆弱性

問題は発見ではなく「パッチ適用の速さ」に移った。

知的財産（IP）と法制度への衝撃

「発明者」概念の空洞化



AIが開発の80%を担う中、現行法の「人間による知的創造」要件維持が困難に。

企業の「AI活用格差」が特許戦略を左右

AIによる解析・出願の速度が人間を圧倒し、FTO分析などの既存プロセスが崩壊するリスク。