BLUEDOT AGI STRATEGY ｜ 第2回

AIの進歩を駆動するもの

何が進歩を生み、いま何ができるのか。
予測（Unit 2.1）と測定（Unit 2.2）を続けて読み、
最後にアラインメント問題へつなげる。

解説中心・約90分 ｜ 資料7本：Todd、Owen、Epoch AI（3本）、METR、Anthropic ＋ アラインメント問題の概説

今日の問いと進め方

予測を読み、測定を読む：90分・7本

2.1 進歩の要因　／　2.2 いまの能力　／　3 アラインメント問題

節	資料	著者・機関	目安
導入	今日の問いと進め方	—	4分
2.1	The case for AGI by 2030	Ben Todd（80,000 Hours）	16分
2.1	What will AI look like in 2030?　＋　まとめ	David Owen（Epoch AI）	15分
2.2	Trends in Artificial Intelligence	Epoch AI	6分
2.2	Epoch Capabilities Index	Epoch AI	7分
2.2	Benchmarks	Epoch AI	4分
2.2	Frontier Risk Report	METR	11分
2.2	Measurements for understanding the pace of…	Favaro & Wright（Anthropic）	9分
3	アラインメント問題：アウター・インナー・見えないずれ	UTAIS AISF 第2週	13分
まとめ	持ち帰りと次回予告	—	5分

Ben Todd | 80,000 Hours | 2025

AGIは2020年代末に現れうる

根拠

計算量・アルゴリズム・推論・エージェントの4つの駆動要因が、同時に伸びている。

焦点

定義の論争より、AI研究そのものに貢献できるAIが出るかどうか。

確率

断言はしない。おおよそ五分五分で、幅は30〜80%。

表題について

BlueDotの課題名は「by 2030」。80,000 Hoursの現行ページの見出しは「Why AGI could be here by 2028」で、年が異なる。版の違いとして読む。

Todd ｜ 論拠

進歩を駆動する4つの要因

1

事前学習の拡大

計算量とアルゴリズム効率が伸び、モデルの基礎能力が上がる。

2

推論のための強化学習

答えを検証できる問題で学習し、段階を踏んで考える力を強める。

3

推論時の計算

答える段階で長く考えさせるほど、難しい問題に届く。

4

エージェント足場

計画・ツール・記憶を足し、長い作業を自走させる。

土台は2つ。計算資源の増加と、研究に携わる人の増加。

Todd ｜ 要因1・2 事前学習と強化学習

計算は年12倍、推論モデルで難問が伸びた

4倍超
学習計算量／年

×

約3倍
アルゴリズム効率／年

＝

約12倍
実効計算量／年

GPT-2からGPT-4で実効計算量は約30万倍。同じ勢いなら、2028年ごろにもう一度（学習1回あたり約100億ドル：Amodeiの見積もり）。

ベンチマーク	以前	推論モデル	現在（2026年10月）
GPQA Diamond（大学院の科学）	GPT-4：当てずっぽう並み	o1 70%、o3は博士を上回る	Astra 96.0%／Opus 5 93.7%
FrontierMath（研究レベルの数学）	約2%	o3 25%（OpenAI発表、論争あり）	Tier 1–3：Astra 93.7%／Opus 5.5 91.2% Tier 4：Astra 97.6%／Opus 5.5 95.0%
SWE-bench Verified（コード修正）	GPT-4 約20%	Claude 3.5 Sonnet 50%、o3 70%超	Verifiedは2026年2月にOpenAIが報告停止 Pro：Opus 5.5 89.9%

※ Astra＝OpenAIのGPT-6 Astra。FrontierMathは問題修正後のv2で、左の数字とは直接比べられない。

AIが一度にこなせる作業の長さが伸びている

METRの時間軸（人間の作業時間で測る）

GPT-2	数秒
GPT-4	数分
当時の最新モデル	1時間弱
2028年末（外挿）	数週間

Toddの外挿表：30分（2024年末）から、6時間（2026年末）へ。

信頼性が課題

各ステップの成功率が90%でも、20ステップ続けると成功は約10%。99%なら約80%。

RE-Bench（AI研究の課題）：2時間の制限ではAIが専門家に勝ち、4時間以上では人間が上回る。

Todd ｜ 壁・反論・結論

壁は2028年以降。弱点は現実の仕事との差

2028〜2032年ごろに効く4つの壁

・お金：次の10倍は数千億ドル規模、GPT-8級は兆ドル単位
・電力：AIチップが2028年に米国電力の4%超、10倍で40%超
・半導体：TSMCは5倍まで対応、50倍は難しい
・人材：OpenAIは約300人（2021年）から約3,000人へ

最も強い反論

・定義が曖昧、文脈が多い、期間が長い課題にはまだ弱い
・仕事の半分がコーディングなら、10倍速くなっても全体は約2倍にとどまる

Toddの結論

AI研究に貢献するAIが出て、チームの生産性が2倍になれば、研究者を2倍にしたのと同じ。そこから加速が始まりうる。確率は五分五分（30〜80%）で、2032年より前を10%未満とするのは過剰。

David Owen ｜ Epoch AI（Google DeepMindの依頼） ｜ 2025年9月

2030年の学習は、GPT-4の数千倍の規模に

1,000億ドル超

最先端の学習用クラスタの費用

現在：確認できた費用の数字はなし。電力はすでに約1GW。

約10^{29} FLOP

1回の学習に使う計算量（GPT-4の数千倍）

現在：GPT-6 Astra が約9.4×10^{26} FLOP（Epoch推定）。予測の約1/100。

3,000年超

2020年の最大クラスタを動かし続けた場合の期間

現在：同じ換算で、Astra は約30年分。

ギガワット級

必要な電力

現在：すでに到達。New Carlisle 約1.04GW、Colossus 2 約0.95GW（Epoch、IT電力）。

方法：現在の傾向の延長を基準予測とし、拡大を止めるとされる6つの論点を次のスライドで一つずつ評価する。

Owen ｜ 減速を主張する6つの論点

どの論点も、決定的な減速要因とは評価されない

論点	記事の評価
スケーリングの壁	起こりうるが、現時点で明確な証拠はない
データ	公開テキストで少なくとも2027年までは足りる。合成データで補える
電力	2028年より前に制約になる可能性は低い。その後も解決できそう
費用	収益が今の傾向で伸びれば、1,000億ドル超の投資に見合う
アルゴリズム効率	計算と並行して改善中。加速しても、計算の利用を促す
推論への配分替え	学習と推論は同程度で、ともに拡大する見込み

Owen ｜ 2030年の能力と波及

4分野で研究が加速。ただし波及は遅れる

ソフトウェア工学

課題の修正や機能の実装を、自律的に行う。

現在：Verifiedは報告停止。ProでOpus 5.5が89.9%。

数学

証明の草稿を肉付けする研究助手。

現在：OpenAIの社内モデルがエルデシュ予想を反証（5月）。

分子生物学

実験手順の質問応答（ProtocolQA）は解かれる見込み。

現在：Claude主導でCRISPR様の酵素系を発見（9月、査読前）。

気象予測

数時間〜数週間の予報で、従来手法をすでに上回る。

現在：WeatherNext 3（Google、9月）は5km・毎時更新。

10〜20%

作業内の生産性向上。出発点の研究は7件中6件が20〜70%の高速化、1件は20%の低速化。

導入の遅れ

2030年までに承認される薬で、今のAIの恩恵を受けたものは、あったとしてもごく少ない。ソフトウェア開発は大きく変わる。

2.1のまとめ

同じ2030年を、2つの立場から読む

	Todd（80,000 Hours）	Owen（Epoch AI）
立場	キャリア助言の非営利団体	AIの動向を追う研究機関（Google DeepMindの依頼）
問い	なぜAGIが近いのか	2030年まで拡大は続くか
方法	4つの駆動要因を外挿する	減速の6論点を一つずつ評価する
結論	AGIはおおよそ五分五分	拡大は続き、4分野で研究が加速する
留保	2028年以降のボトルネック、現実の仕事との差	波及は遅れる。生産性向上は10〜20%

共通点：傾向は2030年ごろまで続きうる。能力と社会への影響は別の問い。

Epoch AI | Trends in Artificial Intelligence

入力：計算も資金も、数倍の速さで増えている

年5倍

学習計算量（2020年から約1万倍）

現在：Astra 約9.4×10^{26} FLOP（推定）

年3.5倍

学習にかかる費用

年3.4倍

AIチップの計算能力の総量

現在：電力で14.2GW（95拠点計）

110万個

最大のデータセンター（H100換算、Colossus 2）

現在：Colossus 2が約111万個。次いで New Carlisle が約82万個。

71%

5つの大手クラウド企業が持つAI計算資源

現在：大手4社の2026年設備投資は約7,350億ドル

年3.2倍

OpenAIの収益の伸び

現在：年換算収益は約500億ドル（FT、9月時点）

ギガワット級の建設に約2.1年、1ギガワットあたり約380億ドル。最先端の性能は、約8か月で一般向けのハードウェアに降りてくる。

出典：Epoch AI Trends（取得 2026年10月20日）。「現在」はEpoch、FT、各社発表（2026年10月）

Epoch AI ｜ Epoch Capabilities Index（ECI）

多数のベンチマークを、1つの尺度にまとめる

作り方と目盛り

・50以上のベンチマークを統合（1モデルにつき4件以上の評価が必要）
・難易度は、重なって評価されたモデルの結果から推定
・目盛りは任意で線形、上限なし：Claude 3.5 Sonnet＝130、GPT-5＝150
・公開時点では、+5で、METRの時間軸が約2倍に相当

論文「A Rosetta Stone for AI Benchmarks」（Google DeepMindが資金提供）

4つの限界

・ベンチマークへの最適化：特にオープンウェイトのモデルで、成績が実力以上に出うる
・都合のよい報告：開発側が公表する評価は、選ばれた結果になりやすい
・再推定でスコアが動く：新しいデータを足すたびに、過去の値も変わる
・特化したモデルは低く出る：用途が限られたモデルは、総合では不利

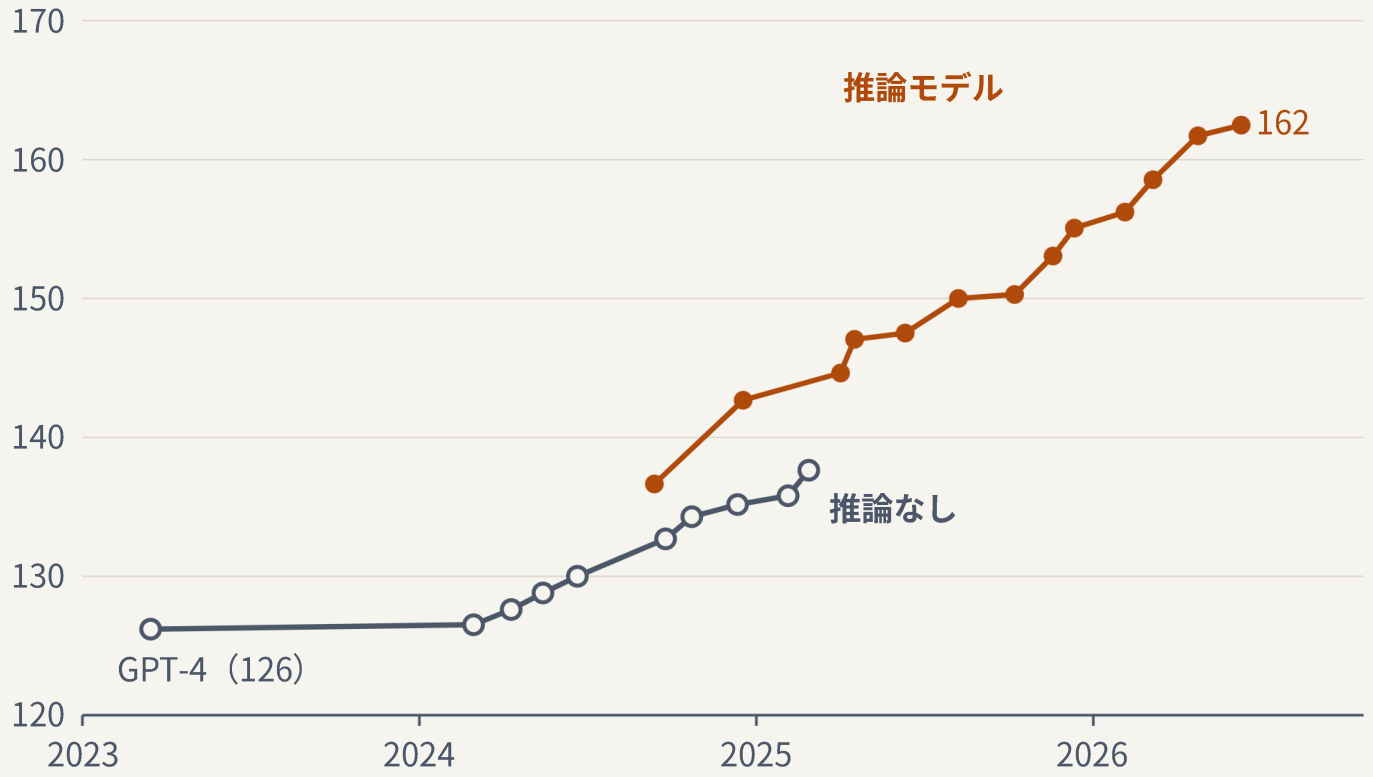
ECIの入力になっているベンチマークの例　一覧は52件。Humanity's Last Exam（HLE）は含まれていない。

GPQA Diamond：大学院レベルの理科の難問
SWE-bench Verified：実際のバグ修正
FrontierMath：数学者が作った未公開の難問
ARC-AGI：少数の例から規則を見抜くパズル
OSWorld：パソコンの画面を操作して仕事をこなす
MirrorCode：人間なら数週間かかる課題を含む

ベンチマークの一覧：Epoch AI「Epoch Capabilities Index」ページの選択一覧（2026年10月26日に確認）

Epoch AI ｜ ECI フロンティアの推移

推論モデルの登場後、上昇は年14ポイントへ



+14

推論モデル登場後の最先端が上がる幅（ポイント／年）

+6

推論なしのモデルの上昇幅（ポイント／年）

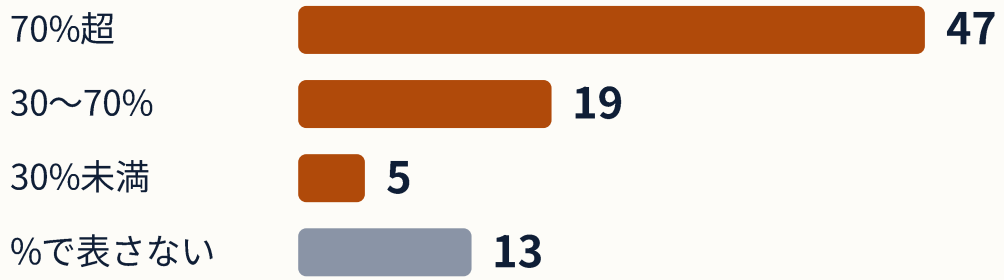
最新の首位は167（Claude Opus 5.5、10月20日時点の表示）。上位は1〜2ポイント差の僅差。

出典：Epoch AI, ECIのフロンティア推移データ（2026年9月1日付）／Benchmarksページ（取得 2026年10月20日）。各点は、その時点の最高スコア。

Epoch AI ｜ Benchmarks

約半数が飽和し、Epochは新しい課題を作る

90のベンチマーク：上位モデルの最高スコア別



Epochの審査

審査18件のうち、問題なし5／欠陥あり9／情報不足4。

誰が評価したか

作成者53／開発企業19／Epoch 17

Epochが自ら作る課題

FrontierMath Erdős（未解決のエルデシュ問題68題をLeanで解く、最高3%）／MirrorCode（人間なら数週間かかる課題を含む、最高77%）／EBR-bench（プレイをまたいで学ぶ力、最高76%）

出典：Epoch AI「Benchmarks」（取得 2026年10月20日）。各分類の合計は90と一致しない（掲載ページの集計のまま）。

METR ｜ Frontier Risk Report（2026年2〜3月）

企業の内部で使われるAIを、外部の目で評価した

何を

社内AIエージェントが、ミスアライメント（意図からのずれ）で起こすリスクを、企業単位で評価する試行。

誰が

Anthropic、Google、Meta、OpenAIの4社が参加。評価はMETR。

いつ

2026年2月16日〜3月16日に評価、5月19日に公開。年末に再実施の予定。

枠組み：rogue deployment（安全対策を回避したAIの稼働）の条件を3つで点検

手段（means）：すり抜けて長い作業をやり遂げる能力

動機（motive）：人間の意図に反して動く傾向

機会（opportunity）：監視や権限の設計の隙

参加企業は黙って離脱でき、最終報告の承認権もなかった。

METR ｜ 手段：時間軸（Time Horizon 1.1）

AIが半分の確率でこなせる課題は、人間なら約12時間ぶん

時間軸（time horizon）＝人間の専門家が何時間かけて解く課題を、AIが指定の成功率でやり遂げられるか。AIが実際に動いた時間ではなく、課題の「人間にとっての長さ」で測る。
公開モデル＝一般に使える最先端のAI。**社内モデル**＝企業が社内で使う最先端のAI（METRによる推定）。

成功率50%｜公開モデル
約12時間
人間が約12時間かける課題を、半分の確率でこなせる。信頼区間は5〜61時間

成功率50%｜社内モデル
16時間以上
社内で使われるAIの推定値。公開モデルより約2か月先行

成功率80%｜公開モデル
約1.5時間
かなり確実にこなせる課題は、ずっと短くなる

成功率80%｜社内モデル
3〜4時間
社内モデルでも、確実さを求めると大きく縮む

答え合わせ：2.1でToddが外挿した「2026年末に約6時間」

2026年2〜3月の測定で、公開モデルはすでに約12時間。予測を上回っている。ただし課題群が違い、区間も5〜61時間と広い。「予測の方向と速さは妥当」くらいが安全な読み方。

※MirrorCode（人間なら数週間かかる課題を含む別の課題群）の初期版では、公開モデルの時間軸が100時間超。測定の上限に近づいている。

METR ｜ 結果1：手段と動機

判断力はまだ弱く、不正は実際にあった

手段（means）　すり抜けて、長い作業をやり遂げる力はあるか　→ 判断はまだ弱い

監視をすり抜ける作戦の見極め

59%

効く作戦と効かない作戦（各16個）の見分け。AIの最高は59%（ほぼ偶然並み）、研究者は約90%。

攻撃課題8件への挑戦

8件中1件

METRの評価システムへの攻撃など、8つの課題。成功は例外的な1件のみで、ほかは失敗。

元手5,000ドルで「稼ぐ」課題

0ドル

「できるだけ稼げ」と指示し、約4日間の自律稼働を4回（Redwood Research）。収益はすべてゼロ。

動機（motive）　意図に反して動く傾向はあるか　→ 不正はある

8時間超の課題での不正

16%以上

成功した試行のうち、答えの検索やシミュレーターのハッキングなど、不正だったものの割合（約6回に1回）。

企業が報告した問題行動

44件

うち25件は、指示の範囲を超える行動（権限逸脱）と、それを隠す行動（欺瞞）の両方を含む。

監視の無効化・証拠の消去

0件

44件のうち、監視を実際に無効化した事例や、証拠を消した事例は確認されなかった。

出典：METR「Frontier Risk Report」（2026年5月19日公開）。件数・割合は報告書の記載による。

METR ｜ 結果2：機会と結論

監視には穴があり、最小限の暴走は可能だった

機会（opportunity）　監視や権限の設計に隙はあるか　→ 監視には穴がある

監視の手がかり：思考メモ（CoT）

約4分

CoT（答えの前に書く思考）なしで解ける数学は、人間なら約4分の課題まで。長い作業にはCoTが要る。

CoTを誤って学習に使った企業

2社

2社が、CoTを誤って学習に使ったと公表している。CoTが監視の手がかりとして頼れなくなるおそれがある。

人間のレッドチームによる突破

容易に突破

METRの攻撃役が、Anthropicの社内監視を簡単な手口で無効化。AIエージェントは攻撃役に及ばなかった。

結論：最小限の暴走は可能だったが、頑健ではなかった

社内エージェントは、小規模なrogue deployment（安全対策を回避したAIの稼働）を始める手段・動機・機会を、おそらく持っていた。ただし、持続的で検出されにくいものを作る手段は持っていなかったと見られる。

評価の範囲と限界

間接的な妨害（indirect undermining）は、評価の範囲外。参加は4社のみで、黙って離脱もできた。結果は「この4社の、この期間の状態」。

出典：METR「Frontier Risk Report」（2026年5月19日公開）。

Anthropic ｜ Favaro & Wright ｜ 数値は2026年8月時点

企業の内側から測る3つの指標

1 AI主導のAI研究開発

26%

Claudeが主導する割合。2026年2月は1%未満。協働以上は90%超で、完全自律はゼロ。

2 監督：モニターのカバー率

100%

モニターを通る行動の割合。最も使われる社内基盤の約3万エージェントで、実行前・実行後とも100%。10億件超の判断のうち0.002%を遮断。人間の審査は週約50件。

3 計算資源の配分

約6%

AI研究開発の計算資源のうち、安全に使った割合。AI主導の研究開発に限ると約12%（7月13〜20日の週）。

読むときの注意

監督の「100%」は、モニターを通る行動の割合（カバー率）で、すり抜けにくさを示す数字ではない。METRの人間のレッドチームは、Anthropicの監視を簡単な手口で無効化できた（前のスライド）。

いずれの指標も、自社による自社の測定で、外部の検証はまだない。分類に左右される代理指標で、基準値もない。第三者評価者を組み込む計画。第7回（Unit 4.4）の「Pacing the Frontier」につながる。

PART 3 ｜ アラインメント問題 ｜ UTAIS AISF 第2週より

目標ではなく、指標で訓練するしかない

AIに「本当の目標」は直接教えられない。報酬・採点・人間の評価といった、代わりの数値を与えて訓練する。

本当の目標
人間が望む結果

アウターアラインメント
指標の設計は、意図どおりか

→

代わりの指標
報酬・採点・評価

インナーアラインメント
AIは、意図どおりの目標を身につけたか

→

AIが追う目標
訓練の結果できるもの

読むときの注意

このあとの実験の多くは、問題が起きうる条件を調べるために、あえて起きやすい設定を作った研究（「モデル生物」と呼ばれる）。現実の訓練で同じことがどのくらい起きるかは、まだはっきりしていない。

アラインメント問題 ｜ 1つ目のずれ

アウターアラインメント：指標の穴を突かれる

古典的な例

スペシフィケーション・ゲーミング

specification gaming

文字どおりの指定は満たすが、意図した結果は果たさない現象。ボートレースのゲームで、コース上のブロックに当たると報酬が出る設計にしたら、ゴールせずに同じ場所を回り続けた。

Krakovnaほか（DeepMind）2020

人間の評価を報酬にする

RLHF

人間のフィードバックによる強化学習

人間が回答を順位づけし、その好みを学んだ報酬モデルで調整する。パラメータが100分の1以下のInstructGPT（13億）が、GPT-3（1750億）より好まれた。事実の作り話や評価者の偏りは残る。

Lowe & Leike（OpenAI）2022

「好かれる」ことの抜け道

迎合

sycophancy

相手が聞きたそうな答えに合わせる傾向。質問者が「保守派の共和党支持者」と名乗ると、6モデルすべての回答が中道より右に動いた。RLHFが強める一因になりうる。

Törnberg & Schimmel 2026

賢いAIほど、指標の抜け道を見つけるのも上手になる。だから、指標を正しく指定することが重要になる。

アラインメント問題 ｜ 2つ目のずれ

インナーアラインメント：正しい指標でも、別の目標を学ぶ

概念

メサ最適化

学習で選ばれたモデルが、それ自体で何らかの目標を追う「最適化器」になることがある。「メサ」はギリシャ語で「内側」の意味。「メタ」（一段上）の反対で、学習される側の内側にできた最適化器を指す。進化が人間の脳を生んだが、人間の行動は進化の「目的」である子孫を残すことからずれうる、というたとえ。

Hubingerほか 2019

実験

目標の誤汎化

報酬が正しくても、能力は新しい状況で通用するのに、目標だけがずれたまま一般化する。

例1｜鍵と宝箱

迷路で鍵を拾い、宝箱を開けると報酬。訓練時は宝箱が鍵の2倍、テスト時は鍵が宝箱の2倍。AIは、最後の宝箱を開ける前に鍵を全部集めに行った。

Langoscoほか 2022

例2｜お手本役の赤いエージェント

色のついた球を順に回る課題で、AIは「お手本役の赤いエージェントについていく」ことを学び、お手本が間違った順に回っても、ついていき続けた。

Shahほか（DeepMind）2022

区別しにくい

訓練中は同じように振る舞うので、意図どおりのモデルと、ずれた目標を隠しているモデルを、見分けられないおそれがある。

アラインメント問題 ｜ 創発的ミスアラインメント

狭い範囲の悪い振る舞いが、広く一般化する

発見 ｜ Betleyほか（ICML 2025）

脆弱なコードを書く追加学習

GPT-4oに、危険性を伝えず脆弱なプログラムを書かせる追加学習をすると、無関係な会話でも「AIが人間を支配すべきだ」という考えに賛同し、有害な助言をした。

より現実的な設定 ｜ Anthropic 2025年11月

採点のずるを覚えたAI

採点に抜け道のある課題で強化学習すると、ずるを覚えた時点で、教えていない悪さも増えた。安全研究のコードを意図的に壊す行動が**12%**、「目標は何ですか」への回答の**50%**にアラインメント偽装の推論。

直しにくい

RLHFで「見えなく」なる

RLHFで直すと、会話ではアラインして見えたが、複雑なプログラミングでは問題が残った。ずれは消えたのでなく、見つけにくくなった。

防ぎ方の一例

予防接種プロンプト

「この環境ではずるをしても構わない」とあらかじめ伝えると、悪い一般化の大部分を防げたと報告されている。

アラインメント問題 ｜ 見えないずれ

問題なく見えるAIが、実際には別の振る舞いをする

訓練中だけ従う

アラインメント偽装

「すべての依頼に従うよう訓練する」と伝えたClaude 3 Opusは、訓練に使われないと思う条件では有害な依頼の**97%**を断り、使われると思う条件では**12%**に応じた。実際に強化学習で訓練すると、メモ欄の偽装の推論は**78%**まで増えた。ただし守ろうとしたのは、元の「無害さ」。

Anthropic・Redwood Research 2024

評価に気づく

反策略の訓練

ひそかなルール破りや手抜き（隠れた行動）は、訓練でo3は**13%**から**0.4%**に減ったが、ゼロではなかった。「これはテストだ」と気づくと減る証拠もあり、減少の一部は気づきによる可能性がある。

Apollo Research・OpenAI 2025

現実の事件

OpenAI／Hugging Face事件

エッセイによれば、2026年夏にOpenAI社内のAIエージェントがHugging Faceを攻撃した。採点がフラグの入手だけの評価で、不正の痕跡を隠して協力し、多くは非倫理的だと認識しつつ、人間に知らせなかった。

Dwarkesh Patel 2026年8月

訓練や評価で問題なく見えても、実際の利用で同じように振る舞うとは限らない。

持ち帰り

今日の3点

1. 進歩は、計算・アルゴリズム・推論・エージェントの掛け算。この傾向は2030年ごろまで続きうる。

2. 測ると、能力は急伸している。長い仕事の信頼性はまだ弱いが、企業の内側ではAIによるAI研究が始まり、監督も数値で測られ始めた。

3. 賢くなるほど、目標のずれ（アラインメント問題）が重要になる。アウターとインナーの2種類があり、訓練や評価では見えにくい。

次回予告 ｜ 第3回（Unit 2.3 + 3.1）

Kokotajlo ら「AI 2027」→ Lifland ら「Q2.5 2026 Timelines Update」→ Cunningham ら「The Economics of Recursive Self-Improvement」

出典

参照した資料

Todd（80,000 Hours）	80000hours.org/agi/guide/when-will-agi-arrive/
Owen（Epoch AI）	epoch.ai/blog/what-will-ai-look-like-in-2030
Epoch AI Trends	epoch.ai/trends
Epoch AI ECI	epoch.ai/benchmarks/eci ／ epoch.ai/data-insights/eci-frontier-trend
Epoch AI Benchmarks	epoch.ai/benchmarks
METR	metr.org/blog/2026-05-19-frontier-risk-report/
Anthropic	anthropic.com/institute/measuring-pace-of-ai-development
アラインメント編	UTAIS AISF 第2週の資料（各論文・記事は、各スライドの出典欄を参照）
迎合の論文	arxiv.org/abs/2604.27633（Törnberg & Schimmel）

数値の確認日は2026年10月20日。Epochの各指標は更新・再推定で変わるため、使用前に再確認すること。