# アンソロピック「When AI builds itself」徹底分析

自己改善 AI と「開発一時停止」提言の検証、および今後の影響

Claude Opus 4.8

2026 年 6 月 6 日

## 要旨

**報道の元になった一次情報**は、Anthropic が 2026 年 6 月 4 日に内部研究組織「The Anthropic Institute」名義で公開したブログ記事「When AI builds itself」（著者：Marina Favaro／Jack Clark、URL スラッグは *recursive-self-improvement*）である [1]。日本語報道はおおむね正確だが、本提言は「世界に開発を遅らせる／一時停止する**“選択肢（option）”**を持たせるべき」という**条件付き・仮定法の任意提言**であり、「即時の開発停止の呼びかけ」と単純化するのは誤りに近い [2]。

Anthropic は「再帰的自己改善（recursive self-improvement, RSI）」に**まだ到達しておらず、不可避でもない**と明言しつつ、社内コードの 80%超を Claude が書いている等の内部データを開示した [1]。提言の核心は「複数国の主要研究所が**検証可能（verifiable）**な形で同時に減速・停止できる仕組み」を事前に構築することであり、一方的・即時の停止要求ではない [3]。

IPO 直前のタイミング（2026 年 6 月 1 日に SEC 秘密申請、評価額約 9,650 億ドル）[8] と、2026 年 2 月の RSP 中核コミットメント後退 [5] から、「規制による競合潰し」「マーケティング」との批判が強い。知財実務では、自己改善 AI＝AI の自律的発明創出が進んでも、現行法上「発明者は自然人に限る」点（日本の知財高裁 2025 年 1 月 30 日 DABUS 判決など）が当面の前提であり続ける [11]。

## 1. 一次情報の特定と提言の正確な読解

### 元記事と中核の提言

元記事は Anthropic 公式サイトの「When AI builds itself」（副題：“Our progress toward recursive self-improvement, and its implications.”）、2026 年 6 月 4 日公開。著者は The Anthropic Institute 代表の Marina Favaro と、共同創業者・政策責任者の Jack Clark である [1]。中核の提言の原文は次のとおり――「社会制度とアラインメント研究が技術の進歩に追いつけるよう、世界が最先端 AI 開発を減速または一時停止する**“選択肢（option）”**を持つことは良いことだと考える」[1]。

**重要な留保**として、原文は「我々は**まだそこに到達しておらず、再帰的自己改善は不可避でもない**。しかし多くの組織が備えているよりも早く来るかもしれない」と明記する [1]。提言自体も「もし効果的にこの技術の開発を減速できるのであれば（If it were possible…）」という**仮定法**で書かれている。

### 開示された内部データ（“証拠”）

- **コード執筆比率：**2026年5月時点でマージされたコードの80%超をClaudeが執筆。Claude Code投入前の2025年2月時点では「数%（low single digits）」だった [1]。
- **生産性：**エンジニア1人あたりの日次コードマージ量は2026年第2四半期に2024年比で**8倍** [1]。
- **外部ベンチマーク（METR）：**AIが自律で確実にこなせるタスクの「時間ホライズン」は約4か月ごとに倍増（従来は約7か月ごと） [1]。
- **タスク到達点：**Claude Opus 3（人間で約4分のタスク）→ Sonnet 3.7（約1.5時間）→ Opus 4.6（12時間タスク）と伸長 [1]。
- **自己留保：**Anthropic自身が「lines of codeは量であり質ではなく、8倍は真の生産性向上をほぼ確実に過大評価している」と認める。社内130名調査（2026年3月）の生産性向上の中央値は約4倍 [1]。

### 提言の正確な「条件」

実効的な停止には、(a) 複数国の、フロンティアにある／近い**複数の資金力ある研究所**が同一条件で停止に合意し、(b) 互いに本当に停止したことを**検証できる**ことが必要だとする [1]。Anthropicは「そうした検証システムが存在すれば、他の最先端開発者も検証可能な形で停止する場合に、当社も減速・一時停止する見込み」と**条件付き**で表明している [3]。一方的停止は「先頭走者が入れ替わるだけで、欠けている熟議プロセスは生まれない」と限界を指摘し、中距離核戦力（INF）全廃条約を緩いモデルとして挙げつつ、検証は核より難しいと認める [1]。

## 2. 報道の正確性の批判的検証

日経・ロイター・ニューズウィーク日本版・Impress Watch等の日本語報道は、原文の趣旨（協調的・検証可能な減速／停止の**“仕組み”**構築の提言）をおおむね正確に伝えている [2]。CNNも「ブレーキペダル（brake pedal）を持つべき」という比喩で本旨を捉えている [2]。

ただし一部の英語見出しには、Anthropicが「即時の世界的開発凍結（global freeze）を呼びか

けた」かのように読める誇張がある [3]。実際は「選択肢を持つべき」という条件付き・将来志向の提言で、原文は仮定法である。もっとも、**「制御不能になる前に開発を遅らせる選択肢を持つことが世界にとって有益」という核心メッセージ自体は、報道は正確に捉えている**。

## 3. 安全枠組み（RSP）との関係と矛盾の指摘

Anthropic の責任あるスケーリング方針（Responsible Scaling Policy, RSP）は、能力閾値（Capability Thresholds）と AI 安全レベル（ASL-1〜）を定め、「自律的 AI 研究開発（Autonomous AI R&D）」を主要閾値の一つとし、到達時には ASL-4 以上の高度なセキュリティと追加の安全保証を要求する設計である。

しかし 2026 年 2 月（RSP v3.0）、複数メディアが「Anthropic が、安全策が事前に十分と保証できない限り AI を訓練しないという**中核的コミットメントを撤回した**」と報じた [5]。最高科学責任者 Jared Kaplan は「（単独で）停止しても誰の役にも立たない」と説明しており、今回の「停止の選択肢」提言との整合性を問う声がある [5]。

## 4. 政策・規制・ガバナンスへの影響

同時期に OpenAI も「フロンティア AI の民主的ガバナンス」ブループリントを公表し、RSI を「緊急の優先課題」と位置づけたとされ、論点設定で主要研究所が足並みを揃えた格好となった。各法域の規制レイヤーの比較は下表のとおり。

| 法域／枠組み | 性格・中核 | RSI への含意 |
|---|---|---|
| EU AI 法 | ハードロー。GPAI で累積訓練計算量 10^25 FLOP 超は「システミックリスク（GPAISR）」と推定され第 55 条の追加義務が課される | 能力ベース閾値の再考を促す可能性 |
| 米国 | 現政権下では比較的「手放し」。政府への安全テスト提出を義務付けない方針 | 州法（SB 53 等）と連邦プリエンプションが論点 |
| 日本 | AI 推進法（2025 年 6 月公布・9 月施行）は罰則なしのソフトロー・推進型。内閣に AI 戦略本部を設置 | 協調的国際停止メカニズムとは方向性が異なり、参加は不透明 |

EU AI 法の GPAISR 義務は、現状は 5〜15 社程度が対象だが数年で対象モデルが急増する見込みで、閾値見直しが論点となる [10]。日本は「世界で最も AI を開発・活用しやすい国」を掲げており [14]、Anthropic の提言する協調的国際停止の枠組みとは方向性が異なる。

### 「Pause AI」公開書簡（2023 年）との比較

2023年3月、Future of Life Instituteが「GPT-4より強力なAIの訓練を少なくとも6か月停止」を呼びかけ3万超の署名を集めたが、誰も停止せず開発は加速した。**差異**は、2023年が外部団体による「即時・一律の6か月停止」要求だったのに対し、2026年のAnthropicはフロンティア研究所自身が内部データに基づき「検証可能な協調的減速の**仕組み**を事前に作る」ことを提言した点にある[4]。両者に共通するのは「単独では止まれない（multipolar trap）」というゲーム理論的問題であり、検証と多国間合意なしには実効性を持たない。

## 5. 産業・市場への影響

Anthropicは2026年6月1日にSECへ秘密裏にS-1を提出。直前のSeries H（650億ドル調達）で評価額**約9,650億ドル**となり、私募評価額でOpenAIを初めて上回ったとされる[8]。年換算収益（run rate）は2026年5月に約470億ドルへ急拡大したと投資家に説明されたと報じられる[9]。提言はこのIPO準備の最中に公表された点が、動機をめぐる議論を呼んでいる。

批判側では、トランプ政権の非公式顧問David Sacksが「規制による囲い込み」と指摘し、複数のアナリスト・学者が「具体策ではなくマーケティング」「一時停止は事実上不可能」と懐疑を表明している[6]。一方、フロンティア研究所がAIによるAI研究加速の内部データを詳細に開示したこと自体が、ガバナンス議論への実質的貢献だとの評価もある。

## 6. 知財（IP）実務への含意

自己改善AIが進み、AIが自律的に発明を生み出す場合でも、現行特許法は「発明者＝自然人（natural person）」を前提とし、各国・地域がほぼ一致してこれを維持している[11]。

**日本の知財高裁DABUS判決：**2025年1月30日判決（令和6年（行コ）第10006号 出願却下処分取消請求控訴事件、控訴棄却）は、「現行特許法は自然人が発明者である発明について特許を受ける権利を認め……AI発明については同法に基づき特許を付与することはできない」と判示した[11]。さらに「AI発明に特許権を付与するか否かは立法化のための議論が必要な問題であって、現行法の解釈論によって対応することは困難」と**立法府に明確に委ねた**。先行する東京地裁判決（令和6年5月16日）も同方向である[12]。

**知的財産推進計画2025：**AI関連発明を4類型（AIモデル発明・AI適用発明・AI利用発明・AI自律発明）に整理し、「発明者の在り方等の諸論点について**早期に結論を得ることが求められる**」とし、検討の場を産業構造審議会・特許制度小委員会と明示した[13]。ただし具体的な期限年は設定しておらず、AI自体を発明者と認める提案には踏み込んでいない[14]。

**実務対応：**知財部門は、AI 支援発明における**人間の創作的寄与の記録**（誰がどのプロンプト・学習データ選択・ファインチューニング・効果確認を行ったか）を整備しておく必要がある。これは米 USPTO の 2024 年 AI 支援発明ガイダンス（自然人が少なくとも 1 つのクレームに有意に寄与すれば特許可能）とも整合する。また、AI の来歴・透明性確保は、Anthropic が提言する compute attestation による停止検証システムと技術的に通底する論点でもある。

## 7. 結論と提言

1. **一次情報を直接参照する（即時）。**報道の二次情報ではなく原文を確認し、条件付き・仮定法の提言である点を正確に押さえる。社内説明では「即時凍結」と誤読しないよう注意喚起する。
2. **「停止の選択肢」と「即時停止」を区別して説明する（即時）。**提言が「複数国・複数研究所による検証可能な協調的減速の仕組みの事前構築」であり、一方的・即時停止ではない点を明示。同時に動機（IPO・競争戦略）を割り引いて評価する。
3. **知財ガバナンスの先回り整備（3〜6 か月）。**AI 支援・AI 生成発明について人間の創作的寄与の記録運用を標準化し、特許制度小委員会の「早期結論」と知財高裁判決の確定状況を継続監視。結論が出次第、社内出願ガイドラインを改訂する。
4. **モニタリング指標（継続）。**(a) 研究所横断の検証プロトコル公表、(b) compute attestation／来歴追跡ツールの採用、(c) 各国の停止・安全テスト義務化の立法、(d) 主要研究所の実際の pause コミット、(e) RSP 能力閾値（Autonomous AI R&D）到達評価——が動けば提言が実体化しつつあるサイン。
5. **規制対応の二層化（6〜12 か月）。**EU AI 法の GPAISR 義務（10^25 FLOP 閾値・第 55 条）への域外適用対応と、日本の AI 推進法・AI 事業者ガイドラインに基づく自主ガバナンス整備を「共通基盤＋各国差分の上書き」方式で並行する。

## 留意事項

- Anthropic の内部データ（80%、8 倍等）は同社の**自己申告**であり外部監査を受けていない。同社自身も指標の限界を明記している。
- IPO 直前という時期、および 2026 年 2 月の RSP 中核コミットメント後退との整合性から、提言の動機には**商業的・競争戦略的側面**がある可能性を割り引く必要がある。

- 「2 年以内に 100%」等は**予測（仮定法）**であり確定事実ではない。Anthropic 自身がトレンドの S 字頭打ち（計算資源・電力制約等）の可能性を第一シナリオとして挙げている。
- 最高裁による DABUS 上告不受理（2026 年 3 月とされる）は単一の二次情報に基づくため、確定情報としての引用前に一次資料での確認を推奨する。
- 本稿は 2026 年 6 月時点の公開情報に基づく。Anthropic は続報（議論の場の設置と結果公表）を予告しており、追跡が必要である。

## 参考文献


[1] Marina Favaro, Jack Clark「When AI builds itself — Our progress toward recursive self-improvement, and its implications」The Anthropic Institute, Anthropic, 2026 年 6 月 4 日（本提言の一次情報）. https://www.anthropic.com/

[2] CNN Business「Anthropic warns that AI will soon be able to improve itself without human intervention」2026 年 6 月 5 日. https://www.cnn.com/2026/06/05/business/anthropic-calls-for-ai-brake-pedal

[3] The Wall Street Journal / To Vima「Anthropic Urges Global Pause in AI Development, Flags 'Self-Improvement' Risk」2026 年 6 月. https://www.tovima.com/wsj/anthropic-urges-global-pause-in-ai-development-flags-self-improvement-risk/

[4] The News「Anthropic's Jack Clark calls for ability to slow down AI progression」2026 年 6 月. https://www.thenews.com.pk/latest/1404873-anthropics-jack-clark-calls-for-ability-to-slow-down-ai-progression

[5] Futurism「Anthropic Drops Its Huge Safety Pledge That Was Supposedly the Whole Point of the Company」（RSP 中核コミットメント後退に関する報道）. https://futurism.com/artificial-intelligence/anthropic-drops-safety-pledge

[6] Digg「（Anthropic の RSI 提言関連まとめ）」2026 年 6 月. https://digg.com/ai/73tec9c0

[7] Sacra「Anthropic revenue, valuation & funding」（評価額・収益データ）. https://sacra.com/c/anthropic/

[8] The Global Statistics「Anthropic IPO Statistics 2026 — Date, Share Price & Valuation」. https://www.theglobalstatistics.com/anthropic-ipo-statistics/

[9] INDmoney「Anthropic IPO Filed Confidentially: Valuation, Revenue, Risks & Investor Guide」. https://www.indmoney.com/blog/us-stocks/anthropic-ipo-valuation-revenue-risks-indian-investors

[10] European Commission「Guidelines on obligations for General-Purpose AI providers」Shaping Europe's digital future（EU AI 法 GPAISR・10^25 FLOP 閾値）. https://digital-strategy.ec.europa.eu/en/faqs/guidelines-obligations-general-purpose-ai-providers

[11] ユアサハラ法律特許事務所「人工知能ダバス（DABUS）に関する令和 7 年 1 月 30 日知財高裁判決（令和 6 年（行コ）第 10006 号）と AI 発明に関する考察」. https://www.yuasa-hara.co.jp/lawinfo/5599/

[12] 長島・大野・常松法律事務所「AI の発明者性について判示した東京地裁判決―東京地判令和 6 年 5 月 16 日―（速報）」. https://www.nagashima.com/publications/publication20240521-1/

[13] 内閣府 知的財産戦略推進事務局「知的財産推進計画 2025 に向けた取組等について」資料 4（産業構造審議会・特許制度小委員会関連）. https://www.meti.go.jp/shingikai/sankoshin/chiteki_zaisan/fusei_kyoso/pdf/026_04_00.pdf

**[14]** 日経クロストレンド「知的財産推進計画や個人情報保護法が AI ビジネスに影響」.
https://xtrend.nikkei.com/atcl/contents/skillup/00009/00162/